

STATISTICAL ISSUES IN SERVICE EVALUATION – A CASE OF INTERMEDIATE CARE

by

BILLINGSLEY CHIMUKA KAAMBWA

A thesis submitted to
The University of Birmingham
For the degree of
DOCTOR OF PHILOSOPHY

Health Economics Unit
Health Services Management Centre
School of Public Policy
The University of Birmingham
December 2008

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

The objective of this thesis was to identify statistical issues that are commonly associated with evaluations of services for older people with a view to establishing the most appropriate methods of addressing them. This goal was achieved in two stages. In the first stage, a comprehensive literature review of studies that have reported such evaluations on populations of older people in the UK was conducted. The second stage involved demonstrating approaches for dealing with these issues on a dataset drawn from largest evaluation of intermediate care done and published in the UK to date. The approaches were adapted from the studies reported in the literature review and where appropriate, from other sources.

This thesis identified a number of statistical issues including those associated with distributional characteristics of variables, missing data and the need to predict utility outcome measures from non-utility ones. Robust approaches of dealing with these problems were demonstrated. The results obtained underlined the importance of avoiding erroneous results and conclusions by applying methods with a sound theoretical background.

ACKNOWLEDGMENTS

It is a pleasure to thank the many people who made this thesis possible.

My supervisors, Professor Stirling Bryan and Dr. Lucinda Billingham deserve special mention. I am grateful to Stirling for having encouraged me to start the PhD in this area and also for helping to firm up my ideas. I also want to sincerely thank him for his calmness, patience and ability to explain things clearly and simply. Lucinda was an inspiration and showed great enthusiasm and an eye for detail in the way only a statistician can. Her great statistical insight and input made my analysis easier. Throughout my thesis-writing period, my supervisors have provided encouragement, sound advice and great direction. Writing this thesis would not have been possible without them.

The data used in this thesis were obtained from the national evaluation of costs and outcomes of intermediate care services for older people in the UK in which I participated. During this evaluation, I collaborated with many colleagues for whom I have great regard and also thank warmly.

I wish to express my gratitude to colleagues at the Health Economics Unit for their encouragement and understanding without which it would have been impossible for me to finish this work. My appreciation goes to my line manager Professor Tracy Roberts for facilitating my 'protected time' to write up the thesis and also to Dr. Pelham Barton for advice rendered during the statistical analysis. For fear of having a

very long list, I will simply say ‘thank you so much’ to the rest of my colleagues for helping in one way or another.

My thanks also go to Dr. Ngenda Mwikisa at the World Health Organisation regional office in Congo-Brazzaville for having inspired and cultivated my initial interest in Health Economics.

I will forever be indebted to my Dad and Mum for instilling a sense of hard work in me from a tender age. They have encouraged me to do the very best in my life and I am truly grateful. And to the rest of my family, I am grateful for your unwavering support in my life.

I owe my loving thanks to my wife Chanda for having endured weekends and evenings filled with PhD work. Your patient love and encouragement has enabled me to complete this work.

And most importantly, to my father in Heaven: I have learnt that that you may not always reveal the purpose of your ways, but your ways are never without purpose.

TABLE OF CONTENTS

CHAPTER ONE - INTRODUCTION

1.0	Introduction.....	1
1.1	Aim and objectives of thesis.....	1
1.2	What is new?.....	2
1.3	Outline of thesis.....	3
1.4	Key messages for health services researchers.....	6
1.5	Definitions of terms.....	8
1.6	Conclusion.....	12

CHAPTER TWO - CARE FOR OLDER PEOPLE

2.1	Introduction.....	13
2.2	Demographic Trends for Older People.....	13
2.3	The needs of older people.....	16
2.4.	The National Service Framework (NSF) for Older People (2001).....	18
2.5.	Development of intermediate care.....	20
2.5.1	Definitions of Intermediate care.....	21
2.5.2	Models of intermediate care.....	23
2.5.3	Developmental milestones of Intermediate Care in the UK since 2000.....	25
2.5.4	Evaluation of Intermediate Care.....	30
2.5.4.1	Search Strategy.....	31
2.5.4.2	How papers were selected for review.....	32
2.5.4.3	Results.....	35
2.5.4.4	Key Findings.....	36

2.5.4.5	Messages from Literature review.....	40
2.5.5	Key messages for health services researchers from this chapter.....	41
2.6	Conclusion.....	41

CHAPTER THREE - LITERATURE REVIEW OF COMMON STATISTICAL PROBLEMS IN EVALUATIONS OF SERVICES FOR OLDER PEOPLE

3.1	Introduction.....	51
3.2	Definitions.....	51
3.3	Summary of Parker et al (2000).....	48
3.3.1	Similarities with the review in this chapter.....	52
3.3.2	Differences from the review in this chapter.....	53
3.4	Methods for the Literature Review	54
3.4.1	Sample.....	54
3.4.2	Search Strategy.....	54
3.4.3	How papers were selected for review.....	55
3.5	Results.....	58
3.5.1	Description of studies from whom data was extracted.....	58
3.5.2	Statistical problems identified in studies.....	61
3.5.2.1	Missing data.....	61
3.5.2.2	Dealing with missing data.....	64
3.5.2.3	Uncertainty of data estimates.....	66
3.5.2.4	Dealing with the problem of uncertainty of data estimates.....	67
3.5.2.5	Lack of generalisability.....	68
3.5.2.6	Dealing with Lack of generalisability.....	69

3.5.2.7	Problems associated with distributional characteristics of variables	70
3.5.2.8	Dealing with problems associated with distributional characteristics of variables	71
3.5.2.9	Problems of sample size and lack of power	72
3.5.2.10	Dealing with problems of sample size and lack of power	73
3.5.2.11	The need for predicted measures of outcome	73
3.5.2.12	Dealing with the need for predicted measures of outcome	74
3.5.2.13	Other statistical issues identified	75
3.5.3	Links between the statistical problems identified in studies	75
3.5.4	Key messages for health services researchers from this chapter	77
3.6	Conclusion	78

CHAPTER FOUR - CASE STUDY: A NATIONAL EVALUATION OF THE COSTS AND OUTCOMES OF INTERMEDIATE CARE SERVICES FOR OLDER PEOPLE – DESCRIPTION OF STUDY

4.1	Introduction	104
4.2	Rationale	104
4.3	Objectives	105
4.4	Methods	106
4.5	Data collected	108
4.6	Baseline Patient Characteristics	112
4.7	Costs	113
4.8	‘Appropriateness’ and intermediate care episodes	113

4.9	Outcomes.....	114
4.10	Identifying statistical problems present in the demonstration dataset.....	114
4.10.1	Missing data.....	115
4.10.2	Lack of generalisability.....	115
4.10.3	Unreliability or inaccuracy of estimates.....	115
4.10.4	Problems associated with distributional characteristics of variables	116
4.10.4.1	Ceiling and floor effects.....	116
4.10.4.2	Too many ‘zero’ values in the dependent Variable.....	116
4.10.4.3	Skewness.....	117
4.10.4.4	Heteroscedasticity	117
4.10.5	Problems of sample size and lack of power.....	123
4.10.6	The need for predicted utility measures of outcome.....	123
4.11	Discussion.....	124
4.11.1	Key findings.....	124
4.11.2	Key messages for health services researchers from this chapter.....	125
4.12	Conclusion.....	126

CHAPTER FIVE - CASE STUDY: A NATIONAL EVALUATION OF THE COSTS AND OUTCOMES OF INTERMEDIATE CARE SERVICES FOR OLDER PEOPLE – STATISTICAL MODELLING

5.1	Introduction.....	133
5.2	Method of analysis.....	133
5.3	Outcome measures.....	134
5.4	Common solutions for dealing with problems associated with distributional characteristics of variables.....	135
5.4.1	Ceiling effects.....	135
5.4.2	Skewness.....	135
5.4.3	Heteroscedasticity.....	136
5.5	Regression Models.....	137
5.5.1	Regression model specification.....	137
5.5.2	CLAD model Specification.....	137
5.5.3	GLM model specification.....	138
5.5.3.1	Choosing an Estimator.....	139
5.5.4	OLS model specification.....	140
5.5.5	Regression Diagnostics.....	141
5.5.6	Hypothesised relationships.....	141
5.5.7	Missing data.....	141
5.6	Results.....	141
5.6.1	Interpreting regression results.....	142
5.6.2	Comparison to OLS results.....	143
5.6.3	Key messages from the analyses.....	145

5.7	Discussion.....	148
5.7.1	Key findings	148
5.7.2	Key messages for health services researchers from this chapter.....	150
5.8	Conclusion.....	151

CHAPTER SIX – CHALLENGES WITH MISSING DATA: THEORY

6.1	Introduction.....	157
6.2	Adverse effects of missing data.....	157
6.3	Mechanisms that lead to missing data.....	158
6.4	Methods for coping with missing data.....	163
6.4.1	Methods for coping with data that are MCAR.....	163
6.4.2	Methods for coping with data that are MAR.....	164
6.4.3	Methods for data that are MNAR.....	170
6.5	Determining the mechanism of missingness.....	171
6.6	Key messages for health services researchers from this chapter.....	174
6.7	Conclusion.....	174

CHAPTER SEVEN – CHALLENGES WITH MISSING DATA: EMPIRICAL ANALYSIS

7.1	Introduction.....	176
7.2	Missing data in the demonstration dataset.....	176
7.3	Reasons for missing data in the demonstration dataset.....	177
7.4	Approaches for dealing with the missing data.....	178

7.4.1	Choice of regression families.....	184
7.5	Results.....	185
7.6	Discussion.....	187
7.6.1	Key findings.....	187
7.6.2	Key messages for health services researchers from this chapter.....	190
7.7	Conclusion.....	191

CHAPTER EIGHT - CHALLENGES WITH OUTCOME DATA: LITERATURE REVIEW

8.1	Introduction.....	201
8.2	Utility versus non-utility measures.....	203
8.3	Previous searches.....	204
8.3.1	Revicki and Kaplan (1993).....	205
8.3.2	Brazier et al (2007).....	205
8.3.3	Mortimer and Segal (2008).....	206
8.4	Search Strategy.....	207
8.5	How papers were selected for review.....	208
8.6	Results.....	210
8.6.1	Description of studies.....	211
8.6.2	Utility measures evaluated.....	211
8.6.3	Non-utility measures evaluated.....	212
8.6.4	Types of analyses.....	212
8.6.5	Key findings.....	214

8.6.6	Methodological issues.....	214
8.6.7	Predicting EQ-5D from Barthel scores	217
8.7	Key messages for health services researchers from this chapter.....	219
8.8	Conclusion.....	220

CHAPTER NINE - CHALLENGES WITH OUTCOME DATA: EMPIRICAL ANALYSIS

9.1	Introduction.....	236
9.2	Methods.....	236
9.2.1	Data.....	236
9.2.2	Statistical Analyses.....	237
9.2.3	Choosing a regression model family.....	242
9.2.4	Robustness.....	246
9.2.5	Predicted EQ-5D indices from models (10) and (11).....	248
9.3	Results.....	251
9.3.1	Demographic characteristics.....	251
9.3.2	Relationship between EQ-5D and Barthel scores–Raw data.....	252
9.3.3	Regression results – Predictive ability of the Barthel.....	258
9.3.4	Predicted EQ-5D Tariff Scores and Goodness-of-fit– Admissions Data.....	268
9.3.5	Robustness - Out-sample Prediction – Discharge Data.....	270
9.3.6	Robustness – Using split Admission Prediction dataset.....	271
9.3.7	Robustness – Using split Admission Validation dataset.....	272
9.3.8	Interpreting the goodness-of-fit results.....	273
9.4	Performance of the models in the validation datasets.....	278
9.4.1	Discrepancies between the prediction and validation datasets.....	278

9.4.2	Similarities between the prediction and validation datasets.....	279
9.4.3	Are the mapping functions stable?.....	280
9.5	Comparison between CLAD and OLS models.....	290
9.5.1	Within-Sample Prediction: Admission Data (Prediction Dataset)....	290
9.5.2	Within-Sample Prediction: Split Admission Data (Prediction Dataset).....	291
9.5.3	Out-of-Sample Prediction: Split Admission Data (Validation Dataset).....	292
9.5.4	OLS or CLAD?	293
9.6	Discussion.....	301
9.6.1	Key findings.....	301
9.6.2	Key messages for health services researchers from this chapter.....	308
9.7	Conclusion.....	310

CHAPTER TEN – DISCUSSION

10.1	Summary of Findings.....	311
10.2	Intermediate care and catering for the needs of older people.....	312
10.3	Common statistical problem in evaluations of services for older people.....	313
10.4	Contribution to literature and key messages to health services researchers from this thesis Contribution to literature.....	320
10.5	Strengths of the research.....	327
10.6	Weaknesses of the research.....	329
10.7	Recommendations for the future	331
10.7.1	Research.....	331
10.7.2	Policy.....	332
10.8	Conclusions.....	332

APPENDIX

A1 - Key terms used to search for literature in Chapter two.....	334
A2 - Key terms used to search for literature in Chapter three.....	335
A3 - Barthel ADL index questionnaire.....	336
A4 - EuroQol EQ-5D (EQ-5D) questionnaire.....	340
A5 - Results from the Heckman selection model.....	342
A6 - Search terms for literature review in chapter eight.....	344
A7 - Regression Results obtained when using the split Admissions Dataset.....	346
A8 – OLS Regression Results obtained when using the Admissions Dataset.....	356
A9 – OLS Regression Results obtained when using the split Admissions Dataset...	363
A10-Plots of Residuals (Admission & Split Admissions datasets).....	370
Kaambwa et al (2008) paper.....	Following page 387
REFERENCES.....	388

LIST OF TABLES

Table 2.1:	Evaluations of Intermediate Care.....	43
Table 3.1:	Studies excluded at Stage IV.....	82
Table 3.2:	Studies selected after Stage IV.....	84
Table 3.3:	Statistical Problems Identified.....	99
Table 4.1:	Summary characteristics of the case-study sites.....	126
Table 4.2:	Variables for use in economic analysis (with level of completeness).....	129
Table 4.3:	Baseline Patient Characteristics.....	130
Table 4.4:	Intermediate care episode costs and health changes (mean, SD).....	131
Table 4.5:	Intermediate care episode costs and health changes (mean, SD).....	132
Table 5.1:	Variables for use in economic analysis.....	152
Table 5.2:	Change in EQ-5D Model – GLM Model.....	153
Table 5.3:	Change in Barthel Model – GLM Model.....	154
Table 5.4:	Change in EQ-5D Model – OLS Model.....	155
Table 5.5:	Change in Barthel Model – OLS Model.....	156
Table 7.1:	Comparison of Results from three methods of Regression Analysis of costs per patient.....	192
Table 7.2:	Comparison of significant variables in the three cost models.....	194
Table 7.3:	Comparison of Results from three methods of Regression Analysis (Change in EQ5D).....	195

Table 7.4:	Comparison of significant variables in the three EQ5D models.....	197
Table 7.5:	Comparison of Results from three methods of Regression Analysis (Change in Barthel).....	198
Table 7.6:	Comparison of significant variables in the three Barthel models	200
Table 8.1:	Studies selected after Stage III.....	222
Table 9.1:	Sample demographic and baseline characteristics.....	252
Table 9.2:	CLAD model – Barthel dimensions entered as continuous variables: Admission Data.....	263
Table 9.3:	CLAD model – Barthel dimensions entered as categorical variables: Admission Data.....	265
Table 9.4:	Multinomial logit using Barthel dimensions as independent variables: Admission Data (Models 10 & 11).....	267
Table 9.5:	Predicted versus Actual EQ-5D utility scores: Admission and Discharge Datasets – CLAD Results.....	275
Table 9.6:	Predicted versus Actual EQ-5D utility scores: Split Admission dataset samples – CLAD Results.....	276
Table 9.7:	Summary of performance of predicted EQ-5D scores ^{\$} - CLAD Results.....	277
Table 9.8	Comparisons of the mean, median, RMSE and MAE for 4 CLAD Models.....	277
Table 9.9:	Predicted versus Actual EQ-5D utility scores: Admission and Discharge Datasets – OLS Results.....	284
Table 9.10:	Predicted versus Actual EQ-5D utility scores: Split Admissions Datasets – OLS Results.....	285
Table 9.11	Differences between OLS and CLAD Predicted EQ-5D utility scores.....	286
Table 9.12:	Comparison between OLS and CLAD results - Within-Sample Prediction: Admission Data (Prediction dataset).....	287
Table 9.13:	Comparison between OLS and CLAD results - Within-Sample Prediction: Split Admission Data (Prediction dataset).....	288

Table 9.14:	Comparison between OLS and CLAD results - Within-Sample Prediction: Split Admission Data (Validation dataset).....	289
Table 9.15:	Comparing items in the Barthel ADL Index and the EQ-5D.....	290
Table A1:	Heckman Selection Models – Costs, EQ-5D and Barthel.....	342
Table A2:	CLAD model – Barthel dimensions entered as continuous variables: Admission Data.....	351
Table A3:	CLAD model – Barthel dimensions entered as categorical variables: Split Admission Data.....	353
Table A4:	Multinomial Logit using Barthel dimensions as independent variables: Split Admission Data (Models 10 & 11).....	355
Table A5:	OLS model – Barthel dimensions entered as continuous variables: Admission Data.....	359
Table A6:	OLS model – Barthel dimensions entered as categorical variables: Admission Data.....	361
Table A7:	OLS model – Barthel dimensions entered as continuous variables: Split Admission Data.....	366
Table A8:	OLS model – Barthel dimensions entered as categorical variables: Split Admission Data.....	368

LIST OF FIGURES

Figure 2.1:	UK Population Indexed on 1901 (100).....	15
Figure 3.1:	Flow Chart of Literature Review.....	81
Figure 4.1:	Change in EQ-5D: Graph of kernel density estimates.....	117
Figure 4.2:	Change in EQ-5D: P-P plot.....	118
Figure 4.3:	Change in Barthel: Graph of kernel density estimates.....	118
Figure 4.4:	Change in Barthel: P-P plot.....	119
Figure 4.5:	Distribution of costs per patient.....	120
Figure 4.6:	Costs per patient: P-P plot.....	120
Figure 4.7:	Plot of residuals against predicted values – EQ-5D.....	122
Figure 4.8:	Plot of residuals against predicted values – Barthel Index.....	122
Figure 4.9:	Plot of residuals against predicted values – Costs per patient.....	123
Figure 8.1:	Flow Chart of Literature Review.....	221
Figure 9.1:	Scatter plot of composite EQ-5D and Barthel scores at admission.....	253
Figure 9.2:	Box-Plot: Barthel & EQ-5D at admission.....	253
Figure 9.3:	Barthel versus EQ-5D Mobility dimension: Admission Data.....	255
Figure 9.4:	Barthel versus EQ-5D Self-Care dimension: Admission Data.....	255
Figure 9.5:	Barthel versus EQ-5D Usual activities dimension: Admission Data.....	256
Figure 9.6:	Barthel score versus EQ-5D Pain dimension; Admission data.....	256
Figure 9.7:	Barthel score versus EQ-5D Anxiety dimension: Admission data.....	257

Figure 9.8:	Plot of Residuals - Model 3 (OLS), Admission dataset.....	299
Figure 9.9:	Plot of Residuals - Model 3 (CLAD), Admission dataset.....	299
Figure 9.10:	Plot of Residuals - Model 4 (OLS), Admission dataset.....	299
Figure 9.11:	Plot of Residuals - Model 4 (CLAD), Admission dataset.....	299
Figure 9.12:	Plot of Residuals - Model 6 (OLS), Admission dataset.....	300
Figure 9.13:	Plot of Residuals - Model 6 (CLAD), Admission dataset.....	300
Figure 9.14:	Plot of Residuals - Model 7 (OLS), Admission dataset.....	300
Figure 9.15:	Plot of Residuals - Model 7 (CLAD), Admission dataset.....	300
Figures A1-A75:	Plots of Residuals (Admission & Split Admissions datasets).....	370

LIST OF BOXES

Box 2.1: The Eight Standards of the National Service Framework (NSF).....19

Box 2.2: Intermediate Care Service Models.....24

Box 9.1: Monte Carlo Method for obtaining predicted EQ-5D indices
from multinomial logistic regression models.....251

CHAPTER ONE - INTRODUCTION

1.0 Introduction

Service evaluations by their very nature are meant to measure the efficiency and effectiveness of an intervention in meeting and supporting some pre-defined goals (Pegram 1999). They give information on whether a service should be continued, improved, expanded or stopped (Rossi et al. 2004). Evaluations therefore inform policy makers of the value or worth of certain services. Evaluations can be quantitative or qualitative in nature. The need for accurate statistical analysis when carrying out a quantitative evaluation of any service can therefore not be overemphasised (Harris et al. 2005). This will in turn ensure that there is accurate and correct interpretation of statistical findings from such an evaluation.

1.1 Aim and objectives of thesis

The main aim of this thesis is to identify key statistical issues that are commonly associated with service evaluations of older people in the United Kingdom (UK) and explore potential solutions. Using a dataset from the national evaluation of costs and outcomes of intermediate care services for older people in the UK (ICNET 2005), this thesis seeks to demonstrate the use of appropriate approaches for coping with these statistical challenges. This dataset is referred to as the ‘demonstration dataset’ in alternate parts of this thesis. Some approaches used to address the statistical problems in the demonstration dataset were identified from the literature review of studies that have evaluated services for older people in the UK. For some statistical problems, no evidence of remedial statistical approaches could be found in the literature review of studies based on populations of older people in the UK. In such instances, appropriate

methods from other populations were utilised. However the question of whether the prevalence or nature of all statistical problems identified in populations of older people is the same as in populations of younger people was not the focus of this thesis and is therefore not addressed.

Thus, the specific objectives of the thesis are to:

- a. identify statistical problems that would form a framework of common statistical issues one has to contend with when evaluating services for older people.
- b. highlight potential biases that result if these issues are not addressed satisfactorily.
- c. consider approaches that can be used to cope with statistical issues in evaluations of services for older people.

1.2 What is new?

This thesis provides the first comprehensive review of statistical issues relating to service evaluations in populations of older people in the UK. This study also utilises data from the largest evaluation of intermediate care (IC) done and published in the UK to date (ICNET 2005). The evaluation, described in greater detail in Chapter four, focussed on costs and health outcomes and produced important results most notably in terms of the need for rigorous patient selection on admission to intermediate care. The data from the national evaluation presented a chance to analyse a unique and large

dataset containing information on 2,253 older people in the UK drawn from a wide spectrum of locations of intermediate care. This is the first time that a number of statistical problems have been addressed from such a dataset and the results obtained will therefore be an important addition to the body of work that looks at studies evaluating services for older people. The results of this exercise will be useful to researchers and policy makers who are involved in evaluations of services for older people and how these impact on service delivery.

1.3 Outline of thesis

The rest of the thesis is organised into nine chapters. **Chapter two** discusses the different forms that care for older people may take. Emphasis is placed on the UK experience. A number of service models that have been used to meet the needs of older people are then explored, with a more detailed discussion of one of these models: intermediate care. This focus on intermediate care was necessary as the demonstration dataset was drawn from an evaluation of intermediate care services in the UK. The chapter presents the results of a literature review of quantitative evaluations of UK intermediate care services before closing with key messages for health services researchers from the chapter.

Chapter three expands this literature review by focussing on evaluations of all services for older people in the UK (intermediate care included) with a view to identifying key statistical issues that are commonly associated with such evaluations and that often lead to bias in statistical analysis. The focus is therefore still on quantitative evaluations. Seven major categories of statistical issues common in the studies reviewed are then presented in order of prevalence. The principal messages for

health services researchers dealing with similar evaluations, drawn from the literature review are also outlined.

In **Chapter four**, a description is presented of the case study, the national evaluation of costs and outcomes of intermediate care services for older people in the UK which was completed in 2005, from which the demonstration dataset used in the statistical analyses in the empirical chapters of this thesis is drawn. A discussion of the rationale behind and objectives of the evaluation is presented. This chapter then discusses the data collection methods before presenting some of the descriptive statistics of the individuals and intermediate care services that were part of the evaluation. Statistical problems present in the demonstration dataset are then identified and the focus is on three of them: problems associated with the distributional characteristics of dependent variables in regression models, missing data and the need for predicted utility measures of outcome. The crucial messages for health services researchers, based on this chapter, are also tendered.

Chapter five demonstrates how one can deal with the problems associated with the distributional characteristics of a dependent variable. This is done using a regression modelling framework. This analysis was part of the quantitative analysis in the national evaluation of costs and outcomes of intermediate care services for older people in the UK and as is shown in this chapter, ‘straightforward’ analyses using simple methods were not possible. The chapter also presents the key implications for health services researchers working in this or related areas

The problem of missing data is tackled in **Chapter six**. This chapter focuses on theoretical underpinnings of three mechanisms that may be responsible for missing data i.e. missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The advantages and disadvantages of methods used to cope with missing data, each making different assumptions about the missingness, are discussed. A summary of the key messages for health services researchers working with populations of older people is also given.

Chapter seven is an empirical chapter and reports the results of using different methods for dealing with missing data on samples from the demonstration dataset. Three methods, each assuming a different missingness mechanism are demonstrated. A discussion of the results and recommendations then follow. The chapter also presents the key implications for health services researchers working in this or related areas

Up to this point in the thesis, the empirical work on outcomes of patients in the national evaluation of intermediate care in the UK considers both the EQ-5D and the Barthel index. Methods for dealing with statistical problems associated with the distributional characteristics of variables and those for missing data were demonstrated on both outcome measures. **Chapters eight** and **nine** consider the question of whether the properties of one measure can allow for it to be used to predict another where it is not possible to collect data on both outcome measures in an evaluation. In **Chapter eight**, the results of a literature review of studies where a utility-based measure has been predicted from a non-utility-based one are presented. A search protocol similar to the one used in chapters two and three was utilised in this

chapter. Key messages for health services researchers concerning the possibility of mapping between non-utility and utility-based outcome measures are also spelt out.

Chapter nine reveals the results of using various regression models to predict the EQ-5D, a utility-based measure of health-related quality of life (HRQoL) from the Barthel index which is a non-utility-based measure of functional independence. The performance of each model was explored in terms of goodness of fit between the observed and predicted values. The robustness of the predictions obtained from the regression models were also tested on external data. The key messages from this mapping exercise are also provided for health services researchers who may be faced with a situation that requires the EQ-5D to be mapped from the Barthel index

In **Chapter ten**, a discussion of the key messages from the thesis and an outline of the policy recommendations and directions for future research is presented.

1.4 Key messages for health services researchers

In the next eight chapters, the implications for health services researchers working with datasets drawn from evaluations of services for older people, or from related areas, are presented in the discussion sections of these chapters. For each chapter, these messages are based on the analyses or information presented in that particular chapter. In Chapter ten a synthesis of all these messages is presented in order to give a ‘take-home’ message from the whole thesis to readers.

To summarise the key messages from this thesis, the first is that based on a review of literature of evaluations of services for older people, there are at least seven major categories of statistical problems that may be present in a dataset drawn from such populations. In addition, various methods for dealing with these problems exist and these should be used or adapted to avoid getting biased or erroneous results. Out of the seven major categories of statistical problems, this thesis looks at three statistical problems as these were the ones identified in the demonstration dataset:

- problems associated with the distributional characteristics of variables
- missing data
- the need for predicted utility measures of outcome

In dealing with problems associated with the distributional characteristics of variables, another key message from this thesis is that it is important to determine the distributional characteristics of variables before conducting any statistical analyses. Problems of heteroscedasticity and skewness can be dealt with by using a method such as a generalised linear model (GLM) regression approach that is robust to these two problems. The GLM is discussed in greater detail in chapter five. Different results and therefore conclusions can be arrived at if methods that ignore these problems are used.

Reasons why data may be missing vary and the methods for dealing with missingness will also depend on the mechanism behind the missing data. Another key message from this thesis is that the choice of method to be used to account for missing data should not be made arbitrarily but should be based on information collected

prospectively about why the data is missing. Where this is not possible, hypothesis testing should inform this choice.

A number of studies have shown that older people are sometimes not able to self-report their health status using utility-based outcome measures. Therefore, a method of predicting health status and health utility in such situations would be very valuable especially when information on another outcome measure has been collected. A key message from this thesis is that it is possible to reasonably predict the EuroQol EQ-5D (a utility-based outcome measure) from the Barthel index (a non-utility based outcome measure). The term utility is defined below in section 1.5 while both outcome measures are described in more detail in chapter four (section 4.5).

1.5 Definitions of terms

Various terms and phrases will be encountered in this thesis which can potentially be interpreted in many different ways. To help the reader understand the context in which these terms and phrases are used, the following definitions should be assumed whenever any of the following terms and/or phrases are encountered:

- Admission avoidance services – Intermediate care services that help prevent unnecessary admissions to hospital or institutional care.
- Attrition – The loss of relevant observations or individuals in a sample after initial definition of the population that is to be included in a study (Matthews et al. 2004). This leads to a gradual reduction in sample size over some period.
- Ceiling effect – the property of a variable which ensures that its values can never exceed a certain value.

- Clinical trial – A type of study design which measures the effects of introducing an intervention. A clinical trial can be controlled (where there is a control or comparison group) or uncontrolled (with no comparison group).
- Comorbidity – The presence of at least two diseases in an individual's health profile.
- Cost effectiveness analysis- A type of economic evaluation where the outcomes are measured in natural units such as life years gained.
- Cost utility analysis - A type of economic evaluation where the outcomes are measured in units which combine both the length and quality of life such as the quality of life years (QALYs) gained.
- Descriptive study – a study designed to describe the occurrence of, or reveal patterns associated with, a specific condition or intervention without an emphasis on pre-specified hypotheses or on effects of variables on a phenomenon.
- Distributional characteristics – Refers to the way various measures of central tendency and dispersion such as the mean, median, mode and standard deviation are related to each other and what effect these have when the whole variable is analysed.
- Economic Evaluation - The comparison of two or more alternative courses of action in terms of both their costs and consequences (Drummond et al. 1997).
- Floor effect – the property of a variable which ensures that its values can never go below a certain value.
- Functional independence – Refers to how easy or difficult it is for individuals to carry out day to day functions in their lives without support from other individuals.
- Geriatric – concerning older people or the characteristics of the aging process.

- Health related Quality of life (HRQoL) – A measure of the quality of life an individual has with a special emphasis on their health.
- Hypothesis test – A test that is used to either confirm or refute a theory on the basis of sample evidence (Gujarati 1995).
- Intermediate care – This is a service that prevents admission to acute care or long term care and also aids discharge from hospital for older people.
- Missing data – this refers to either observations that are missing or absent within a variable or variables missing in a dataset.
- Non-experimental analytic study – A study that examines associations or hypothesised causal relationships but does not involve a trial-based intervention.
- Older people – This refers to people aged 65 years or over.
- Outcome measure – A way of assessing or ascertaining the result or effect of a certain intervention on an individual.
- Prediction error – The error that is an indicator of the difference between the predicted values and the observed values of a variable of interest. This error can be measured, among others, by the root mean squared error and mean absolute error discussed in more detail in chapter nine.
- Predictive regression model – A regression model that is built to forecast or predict the probability of an outcome.
- Proxy measure – a measure that can be used in the place of another that may be missing or not available for a variety of reasons.
- Psychometric measure – one where a respondent indicates “...the presence, frequency, or intensity of symptoms, behaviours, capabilities or feeling. Responses to individual questions are aggregated to create individual

homogeneous scales (e.g., physical function, social function, mental health) or global summary scales.” (Revicki & Kaplan, 1993, p.477).

- Quantitative analysis – An analysis that focuses on numerical and measurable characteristics of variables such as costs per patient.
- Regression analysis – An analysis where the effect that a change in one variable has on another is measured.
- Skewness – The asymmetry in the distribution of the sample data values (Altman 1991).
- Statistical problem – An analytical problem that arises from the collection, organisation, analysis, interpretation and presentation of data or numbers (Daniel & Terrell 1995).
- Supported discharge services - Intermediate care services that aid patient discharge following a stay in an acute hospital.
- Systematic Review – A method that uses “...explicit and rigorous methods to identify, critically appraise, and synthesize relevant studies.” (Mulrow et al, 1997, p.390).
- Utility – This is the cardinal value that represents the strength of an individual’s preferences for specific outcomes under conditions of uncertainty (Torrance 1986; Torrance & Feeny 1989). It may also be seen as a “...psychological concept of welfare or well-being” (Richardson, 1994, p.8) or based on a standard gamble, individuals’ indication of “...probability that leaves them indifferent between the state to be evaluated and a gamble in which the outcomes are two reference states.” (Richardson, 1994, p.9). Lastly, utility can also be seen as a measure of individuals’ sacrifice of one thing of value to them in order to gain another thing

they also value in such a way that they are indifferent between the two states of the world (Dolan et al. 1996).

1.5 Conclusion

This chapter has outlined the objectives of this thesis and has also highlighted what is original about the work in the thesis. This is the first time that a comprehensive review of statistical issues relating to service evaluation in populations of older people in the UK has been conducted. Further, the dataset that is used to demonstrate approaches to address some of the statistical issues identified in the review is obtained from the largest evaluation of intermediate care done and published in the UK to date. This chapter has also given a flavour of the key messages to health researchers dealing with datasets drawn from populations of older people (or similar populations) that can be obtained from the empirical chapters of this thesis. The next chapter focuses on the different forms that care of older people may take.

CHAPTER TWO - CARE FOR OLDER PEOPLE

2.1 Introduction

This chapter considers the different forms that care for older people takes in many countries, principally the UK. The challenge of caring for older people is first explored by examining the growing needs of older people due to the demographic trends in the world population. Eight approaches or standards for catering to the needs of older people as encapsulated in the UK National Service Framework (NSF) for older people of 2001 (Department of Health 2001b) are discussed. Particular consideration is then given to intermediate care which is one of these approaches. A narration of the developmental phases of intermediate care is provided as well as a summary of studies that have evaluated intermediate care services in the UK. The chapter closes with key messages to health services researchers dealing with populations of older people. This chapter aids understanding of the service models that are available to meet the needs of older people in the UK. A special focus on intermediate care is helpful in understanding this service model especially as the empirical results reported in chapters five, seven and nine are based on a dataset drawn from an evaluation of intermediate care services in the UK (ICNET 2005).

2.2 Demographic Trends for older people

Worldwide

One of the main features of the world population today is the considerable increase in the absolute and relative numbers of older people. The proportion of older people is getting bigger. Estimates show that the world's population will consist of more than two billion people aged 60 years and older by 2050 (United Nations 2006). Another

trend is evident: the older population is ageing also - the 'oldest old', those over 75 years old, are the fastest growing group (Tomassini 2005). It is also expected that the number of people worldwide who will reach 80 years or more will exceed 402 million by 2050. There are many reasons for this trend and one of them is that longer lives have been coupled with falling birth rates resulting in a slower overall population growth but a bigger proportion of older people (United Nations 2006). In many countries, women make up the biggest proportion of older people. Between the ages of 65 and 79, there are now approximately three women for every two men. Women outnumber the men by a factor of two for individuals over the age of 80 (European Institute of Women's Health 1996). On average, estimates show that women live more than six years longer than men. In the year 1999, the highest life expectancy at birth for women was for France (83.6 years) followed closely by Italy and Spain (United Nations 2000).

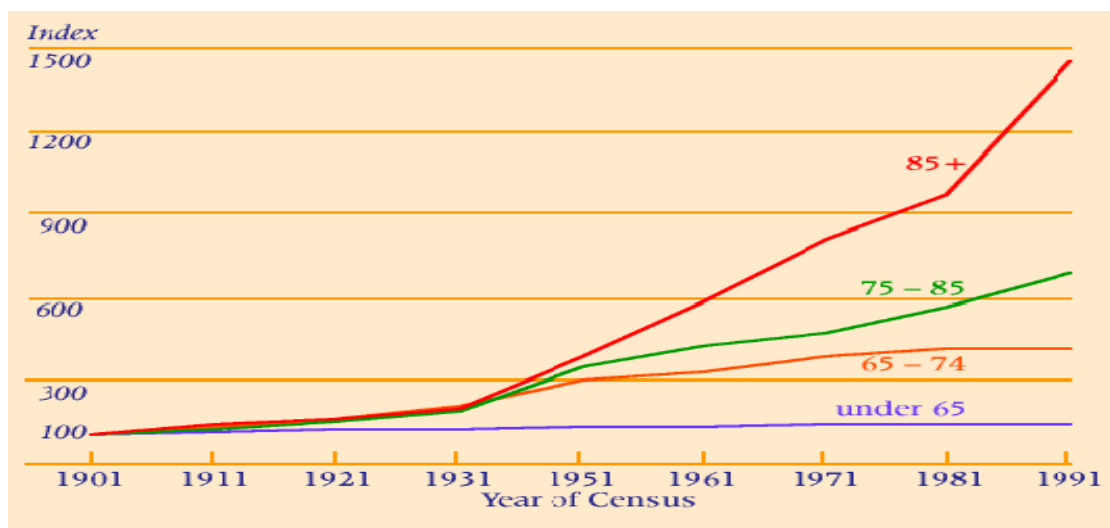
OECD region

An examination of the countries in the Organization for Economic Co-operation and Development (OECD) area - Western Europe, North America and the Asia-Pacific region reveals similar trends in the proportion of the over 65s. It was projected that between 1950 and 2050, this population will have experienced a more than 50 per cent increase (Department of Health 1999). Western European countries are projected to experience an increase of at least 200 per cent in the over 80s during the period 1960/61-2040/41. It is also expected that the growth in non-European industrialized countries will be about 500 per cent in New Zealand, over 800 per cent in the United States, over 900 per cent in Australia and Canada and over 1,300 per cent in Japan (Department of Health 1999).

United Kingdom

The UK population trends have been characterised by unprecedented changes in life expectancy, consequently increasing the proportion of those aged 65 years and over. The beginning of the 20th century was punctuated by a 400 per cent increase in the number of older people, which has doubled since the early 1930's (Royal Commission on Long Term Care for the Elderly 1999) (Fig.2.1).

Figure 2.1: UK Population Indexed on 1901 (100)



Source: Royal Commission on Long Term Care for the Elderly (1999)

Here too, there has also been the secondary ageing process: the ageing of the aged. The result has been an increase in the number of people aged 80 and over (Department of Health 1999). It is projected that by 2050, the number of people aged 80 and over will be three times more numerous (Royal Commission on Long Term Care for the Elderly 1999). The Wanless report of 2002 states that in the next 20 years, those aged 85 and over in places like England are projected to increase by two-thirds, in comparison to a 10 per cent growth in the overall population (HM Treasury 2002).

2.3 The needs of older people

While a number of the needs of older people are the same as those for other age groups, there are some that are specific to the former. According to Clough et al (2007) older people need help around the house and garden (such as general cleaning, laundry and other household chores); staying in and going out (e.g. getting in or out of the bath, feeling safe in the street or neighbourhood); managing personal affairs (such as writing to utilities and others); shopping; transport; socialising; leisure and recreation. Older people also utilise more medical care than other groups (Department of Health 2002b). In the UK, for instance, evidence from the National Beds Enquiry shows that older people occupy two-thirds of general and acute hospital beds, and account for over half of the recent growth in emergency admissions (Department of Health 2002d). Length of stay in hospital is also significantly greater for older people (Department of Health 2002d). There is then always a question of how to best cater for older people's needs. This is because of the heterogeneity in the characteristics of older people and therefore their requirements. Between 1981 and 2001, increases in healthy life expectancy did not keep pace with improvements in total life expectancy. In future, the total number of people with disabilities, and potentially in need of care, will be higher. The NSF (Department of Health 2001b) identifies three broad groups of older people:

a) Those entering old age having completed their career in paid employment and/or child rearing. As this is a socially-constructed definition of old age, it is subject to different interpretations. It may include people as young as 50 years, or from the official retirement ages of 60 years for women and 65 years for men. These groups of

people are characterised by an active and independent life and many remain so into late old age.

b) There are those in a transitional phase between a healthy, active life and frailty. This transition is often associated with those in the 70s and 80s despite the fact that it can occur at any stage of older age.

c) The last group is that of frail older people. These people are associated with health problems such as stroke or dementia and may also have social care needs or a combination of both and are therefore a vulnerable group. Frailty is often experienced in late old age, so services for older people should be designed with their needs in mind.

The need to care for older people can not be disputed. According to Wise (1997), older people are working less and saving less which has made it even more imperative for some third party to be involved in caring for them. This increases the demand for all forms of care, putting pressure on available resources and funding. Further, older people may be suffering from particular conditions or indeed cormobidities of conditions. Common conditions associated with older people include Alzheimer's disease, Arthritis, Dementia, depression, disability, falls, functional dependence, heart disease and stroke, poor health related quality of life and sensory problems (vision and hearing).

Decisions therefore have to be made concerning the appropriate response to these needs or conditions. Each circumstance will require a unique response. This will

initially be a decision about whether what should be given is housing, health or social care social or a combination of any of these. The overriding goal would be to provide the appropriate support and care services that will help older people to remain independent (Department of Health 2001b). The setting in which this care is provided is another variable. Housing can be made available in a number of places, provided the housing units are well-maintained, warm and secure so as to help prevent unnecessary admissions to hospital or institutional care. Health care for the aged can be provided in hospitals, residential or nursing homes as well as in their homes. Social care has been rendered in older people's homes or communities mainly by local authorities (Roe 2005; Department of Health 2001a; Department of Health 2001b).

2.4. The National Service Framework (NSF) for Older People (2001)

This NSF was first published on 2 March 2001 with an overriding aim of addressing the health and social care needs of older people. It is bedded on partnership-working between several players: those who use and those who provide services; between different clinicians and practitioners; across different parts of the UK National Health Service (NHS); between the NHS and local government; between the public, voluntary and private sectors; and individuals, groups and organisations within the community. The NSF was developed with the advice of an External Reference Group and the Social Services Inspectorate. This framework has been rendered even more important considering that older people are more likely to become seriously ill and potentially die than younger people. The framework highlights eight standards that should be met in the care of older people (Department of Health 2001b). These are presented in Box 2.1:

Box 2.1: The Eight Standards of the National Service Framework (NSF)

1: Rooting out age discrimination - It emphasised the need to combat widespread infringement of dignity and the unfair discrimination in older people's access to care on the basis of age. Provision of NHS services would on the basis of clinical need and ability to benefit alone as opposed to the patient's age.

2: Providing person-centred care - Individuals would make choices about their own care. Social care and NHS services would also be centred on the needs of older people. Assessment would be via a single assessment process.

3: Intermediate care - A new layer of care that would help prevent unnecessary hospital admission, support early discharge through the provision of effective rehabilitation services and reduce or delay the need for premature or unnecessary admission to long-term residential care was promoted.

4: General hospital care - Only appropriate specialist carers and hospital staff with the right set of skills should be involved in the delivery of older people's care in hospital.

5: Stroke – Reducing stroke by ensuring that stroke patients had quick access to integrated stroke care services provided by the NHS and other agencies. This implied proper access to diagnostics services, appropriate treatment by specialist stroke services and benefiting from a multidisciplinary programme of secondary prevention and rehabilitation.

6: Falls - The aim was to reduce the number of falls and ensure treatment and rehabilitation of those who had fallen. The NHS was to work with other partners (councils).

7: Mental health in older people - Older people with dementia and depression were to be treated and supported by ensuring that they had access to integrated mental health services provided by the NHS and councils.

8: The promotion of health and active life in older age - In order to lengthen the healthy life expectancy of older people, the health and well-being of older people was to be promoted through a coordinated programme of action led by the NHS and supported by councils.

2.5. Development of intermediate care

Intermediate care, which is one of the eight NSF standards, has undergone many phases of development. As a report that evaluated intermediate care nationally reported, intermediate care is not a new idea (ICNET 2005). It could be argued that community hospitals, community nursing inputs and community-based therapists have promoted independence and have prevented admission to and facilitated discharge from hospital in the last number of years. Nevertheless, there has been an increase in specially designed, usually multi-disciplinary, models of care targeted specifically at achieving early discharge or, more recently, avoiding hospital admission at all (Parker et al. 1999). Intermediate care was first postulated as a formal policy in the *NHS Plan* (Department of Health 2000c) which formed a key component of the NSF for older people (Department of Health 2001b). One of the key drivers for the establishment of intermediate care was the need to connect acute hospital and primary and community care (Department of Health 2001b). There was lack of investment in preventative or rehabilitation services for older people and as a result, this group of people were being unnecessarily admitted to hospital and remaining there longer than was necessary (Audit Commission 1997). There was need therefore to identify alternatives to the expensive use of acute care and premature admission to residential and nursing home care (Audit Commission 1997). Beech et al (2004) submit that this is in line with the 10-year plan for investment in, and

modernisation of, the NHS which would result in an expansion of the capacity of acute hospitals as well as development of substitute services. Further, the National Beds Inquiry (Department of Health 2000b) pointed to the shortages of community-based alternatives to hospital care as well as the significant delayed discharges as evidence of the fact that health and social care systems were not meeting the needs of older people properly. The need for alternative models of care was further highlighted by the fact that about 20% of the bed days in acute hospitals were likely inappropriate (Department of Health 2000a). Intermediate care was therefore seen as a model that would help to release space in acute wards and also reduce waiting times (Department of Health 2000c).

2.5.1 Definitions of Intermediate care

Many settings providing a service to older people have put themselves under the umbrella of intermediate care. But as many commentators have observed, what really constitutes intermediate care remains a subject of debate. Roe et al (2003) and Beech et al (2004) point out that there has been a proliferation of intermediate care schemes since the early 1990s where new schemes have been added onto the existing ones without a clear indication of how they fit as intermediate care services.

In early attempts to define intermediate care, the King's Fund defined it as a 'function' as opposed to a discrete set of services with the concepts of transition and restoration central to it (Steiner 1997a; Steiner & Vaughan 1997). In the first definition, intermediate care was seen as a "... whole set of services designed to smooth transitions between hospitals and home, treat chronically or terminally ill people without recourse to hospital care, prevent long-term institutionalization"

(Steiner, 1997b, p.24). This was later narrowed down to “...that range of services designed to facilitate the transition from hospital to home, and from medical dependence to functional independence, where the objectives of care are not primarily medical, the patient’s discharge destination is anticipated, and the clinical outcome of recovery or restoration to health is desired.” (Steiner, 1997b, p.24). While both definitions include care given to patients with a variety of conditions, in a range of venues, the second definition excludes palliation (British Geriatrics Society 2005). The Audit commission (2000, p.21) further saw the primary function of intermediate care as being that of services that “...provide rehabilitation to people who are medically stable, but who are not yet ready to return home after their discharge from hospital. They can also be used as ‘step-up’ facilities for people living at home who need a period of intensive rehabilitation, but who do not need the full range of inpatient services with specialist medical and nursing support on site.”

The UK Department of Health issued a standard definition of intermediate care as part of their guidance (Guidance HSC/LAC 2001/01) which would ensure that there was a consistent and clear approach to the development, monitoring and benchmarking of intermediate care services (Department of Health 2001a). Five criteria were outlined (ICNET, 2005, p.3):

- They are targeted at people who would otherwise face unnecessary prolonged hospital stays, or inappropriate admission to acute in-patient care, long term residential care, or continuing NHS in-patient care.

- They are provided on the basis of a comprehensive assessment, resulting in a structured individual care plan that involves active therapy, treatment or opportunity for recovery.
- They have a planned outcome of maximising independence and typically enabling patients/users to resume living at home.
- They are time-limited, normally no longer than six weeks and frequently as little as 1-2 weeks or less; and
- They involve cross-professional working, with a single assessment framework, single professional records and shared protocols.

In other parts of the UK such as Scotland, the label of ‘intermediate care’ has not been used with the Scottish Executive preferring to use ‘integrated services’ to encompass ‘good, patient-centred mainstream services’ for older people (Petch 2003).

2.5.2 Models of intermediate care

Intermediate care can be summarized as a variety of services that serve as a bridge between institutionalisation (e.g. in a hospital) and home for people in need of support that requires the blend of health and social care (Department of Health 2001a;Steiner 2001;Stevenson & Spencer 2002). Older people are the main but not exclusive beneficiaries of these services (Department of Health 2001b).

The Department of Health comments that development of intermediate care has resulted in a wide diversity of models where ‘a thousand flowers have bloomed’ (Department of Health 2002b).

While the location is clearly important, the term intermediate may also refer to care organized and delivered by teams of different professionals and organizations (Stevenson & Spencer 2002). Vaughan and Lathlean (1999) submit that the gradual elimination of barriers between doctors and other clinical professionals, social and health services, statutory and non-statutory services is necessary to provide important opportunities that will smooth the many interfaces throughout the system. Box 2.2 below illustrates some of the intermediate care service models that have been followed in the UK (ICNET, 2005, p.4).

Box 2.2: Intermediate Care Service Models

Rapid response – designed to prevent avoidable admissions by providing rapid assessment/diagnosis and rapid access on a 24-hour basis to short-term nursing/therapy support and personal care in the patient’s own home.

Hospital at home – intensive support in the patient’s own home as a way of avoiding an acute admission or to enable earlier discharge from hospital.

Residential rehabilitation – a short-term programme of therapy and enablement in a residential setting (such as a community hospital, rehabilitation centre, nursing/ or residential care home) for people who are medically stable but require a short period of rehabilitation in order to return safely to their own home. Services may be ‘step down,’ following a stay in an acute hospital or ‘step up’, following a community referral and full assessment.

Supported discharge – a short-term period of nursing and/or therapeutic support in a patient's own home, typically including a package of home care support. Sometimes supported by community equipment and/or housing-based support services.

Day rehabilitation – a short-term programme of therapeutic support, provided at a day hospital or day centre. May be used in conjunction with other forms of intermediate care.

Adopted from HSC/LAC 2001/01 (Department of Health 2001a)

Other models or schemes have been identified by Roe (2005) including community rehabilitation teams, community hospitals, sheltered accommodation and assisted living, nurse-led units, care homes and equipment services.

2.5.3 Developmental milestones of Intermediate Care in the UK since 2000

The growth of intermediate care has not seen a consistent approach across the country which has resulted in ‘...confusion and fragmentation, in turn resulting in inequality of provision and access, duplication of effort, reduced cost effectiveness, and loss of impact’ (Department of Health, 2002c, p.5). Whereas the roots of intermediate care can be traced as far back as the 1990s, it did not become incorporated in mainstream Department of Health policy until after 2000. For this reason the emphasis in this section is on the period from 2000 onwards. As outlined in Cowpe (2005), there were some distinctive phases of intermediate care policy development in the UK:

Shaping the Future NHS: Long-term Planning for Hospitals and Related Services.

The National Beds Inquiry (Department of Health 2000b).

One of the key results of this enquiry was the evidence that there was significant inappropriate or avoidable use of acute hospital beds which resulted in about 20% of bed days for older people being potentially inappropriate if alternative facilities had

been available (Department of Health 2000b). As a result, a decision was made to provide care 'closer to home' where there would be an expansion of community health and social care in aid of the development of intermediate care services. The objective of these services was to avoid unnecessary admissions to acute care as well as to aid earlier discharge from hospital and a return to functional independence. The focus of acute hospital services would be on rapid assessment, stabilisation and treatment.

The NHS Plan: A Plan for Investment, A Plan for Reform (Department of Health 2000c).

This plan was issued in July of 2000 and outlined a major new programme that was targeted at promoting the independence of older people through a number of intermediate care and related services. It emphasised three issues: (i) aiding people's recovery and quicker regaining of independence (ii) facilitating easier discharge from acute care and (iii) preventing unnecessary long-term care. Also embedded in the plan was an explicit mention of the provision of an extra £900m investment by the year 2003/04 to support these programmes with the NHS receiving about £405m of this investment. The rest was earmarked for local government. Cooperation was expected between the health and social services while the monitoring of the programme was to be done by the Commission for Health Improvement, the Audit Commission and the Social Services Inspectorate.

Intermediate Care Health and Local Authority Circular HSC 2001/01: LAC (2001)1 (Department of Health 2001a).

This circular is seen as the first detailed statement that outlined government policy on intermediate care. Intermediate care services were regarded as any services that met the following criteria: (i) prevented avoidable admission to health or social care facilities and aided earlier discharge from hospital (ii) had a wide-ranging plan for comprehensive assessment and individual care whose goal was to maximise independence (iii) care was not to last for more than six weeks and (iv) there was to be multi-agency input even though the same assessment framework, record and protocol were to be used. The circular also brought to the fore the description of intermediate care models described in Box 2.2.

National Service Framework for Older People (Department of Health 2001b).

Details of the eight standards of this framework have been presented in Box 2.1. Its aim was to aid a quicker recovery from illness, prevent needless admission to acute hospitals, foster timely discharges from hospitals as well as to optimise independent living through the provision of 'integrated services'. The integrated services would include primary and secondary health care, social care as well as statutory and independent sectors. At least three tenets were seen to be important in ensuring that intermediate care was delivered successfully (Cowpe 2005): (i) health and social care agencies were to be involved in open and effective partnerships that would progress to planning and investment. (ii) access to specialist assessment, diagnosis and treatment was to be ensured for intermediate care patients if needed (iii) the team to provide intermediate care was to be a coordinated one drawing upon a wide range of professionals such as general practitioners, hospital doctors, nurses, physiotherapists,

occupational therapists, speech and language therapists, social workers, care assistants and administrative staff.

Securing our Future Health: Taking a Long-term View. The first Wanless Report (HM Treasury 2002).

One of the conclusions arrived at by this report was that a significantly larger share of the UK's national income must be devoted to health care over the following 20 years. This was seen as a way of reversing the significant cumulative underinvestment over past decades so as to catch up with the standards of care seen in other countries (HM Treasury 2002). This advice is credited to have led to an increase of £40 billion for the health service (up to the year 2007/08) to be announced in the April 2002 budget (Cowpe 2005). This budget also included a 6% increase in personal social services investment.

Delivering the NHS Plan: Next Steps on Investment, Next Steps on Reform (Department of Health 2002a).

This plan emphasised the need for health and social services to work together for patients, especially older people. Only good integration between the two would ensure the delivery of the care older people need, when they need it. The report reaffirmed the gains that had been obtained due to the implementation of the NSF for older people such as the establishment of the new single assessment process, carrying out an audit of age discrimination and heavy investment in intermediate care. Other gains were the local pooling of health and social services budgets through the increased use of Health Act flexibilities as well as increased resources for social services. These gains were viewed as further reform strides towards the NHS vision of a single care

system where the interests of its users, not providers, are central. There was an admission, however, that despite these achievements that broke the 'Berlin Wall' between health and social care, older people in some parts of the country failed to get holistic services required because of the failure of health and social care to cooperate. Proposals for new arrangements for personal social services were announced such as the introduction of a new £50m performance fund for intermediate care for the first time. There was also need to revisit the top-down targeting of resources, central intervention and close monitoring of delayed hospital discharge which would not be sustainable in the long run. There was a determination to introduce new arrangements, in consultation with local government, which would ensure a more seamless service for patients.

***National Service Framework for Older People – Supporting Implementation.
Intermediate Care: Moving Forward (Department of Health 2002c).***

The framework was published to support the NSF for older people, and presented a review of progress in the implementation of intermediate care and pointed out priorities for future development (Department of Health 2002c). It identified 'success factors' based on good practice. The framework pointed out that while intermediate care had made rapid progress over the previous two years, it still was relatively new and faced testing challenges before becoming firmly established in mainstream care. While stressing the importance of the principles of 'patient-centred care' and 'whole-systems working', the document also posited that to realise the potential of intermediate care for both individual service users and health and social care systems, there was need for adherence to these principles. The success of local intermediate care schemes was acknowledged but there was also recognition that inconsistency,

fragmentation, lack of coherence and poor integration with other services in some areas had typified the early stages of intermediate care development. A 'whole-systems' approach for intermediate care was postulated. Further, the role of intermediate care co-ordinators in bringing consolidation, cohesion and consistency to existing intermediate care services was reinforced in the document. In addition, constituting a single point of contact for intermediate care was hailed a 'success factor' in subsequent guidance on hospital discharge (Department of Health 2003). Other future priorities included: (i) the need for making sure that people with mental health problems, including cognitive impairment, have access to intermediate care services (ii) the development of a clinical governance framework (iii) determination of an appropriate level of medical assessment and support (iv) recognition of the role of housing in promoting independence.

2.5.4 Evaluations of Intermediate Care

As already pointed out, Intermediate care in the UK has been in existence in many forms for a number of years. Inevitably, many evaluatory studies have been commissioned to consider evidence on the performance of intermediate care schemes. The department of health recommends that evaluation should be made implicit within the everyday provision of intermediate care services (Department of Health 2001a; Department of Health 2001b). Evaluation should be utilised as a tool that will give guidance to the initial evaluation of intermediate care schemes, to measure success of current services as well as to appraise the long term outcomes of these schemes (Beech et al. 2004; Roe 2005). Pre-2000 systematic reviews of intermediate care services such as Parker et al (2000) concluded that there was lack of evidence on cost-effectiveness, let alone the effectiveness, of intermediate care. Beech et al (2004)

also point out that the extent to which intermediate care services provide a suitable and effective substitute to acute based care is uncertain. In this chapter, a literature review was conducted to identify studies that have evaluated intermediate care services in the UK and focused on the post-2000 period. In this search, only services that have been explicitly named as ‘intermediate care’ or as another service that is widely known to be part of intermediate care (e.g. admission avoidance and supported discharge schemes) have been included. A full list of search terms is provided in A1 of the appendix.

2.5.4.1 Search Strategy

A literature review of notable evaluations of intermediate care services in the UK was undertaken. Using the protocol employed by Mugford (2001), a formal scoping search was undertaken. Key terms (words and/or phrases) were typed in WebSPIRS search boxes and available databases that provided matches were then selected accordingly. The search of databases carried out involved combining the terms shown in the appendix and searching for material from 2000 onwards. A range of online computer databases was searched to identify relevant service evaluation literature:

- ASSIA(CSA): 2000 – 2008
- CINAHL (EBSCO): January 2000 to August 2008
- Cochrane library (Wiley): 2000 - 2008
- EMBASE (Ovid): 2000 – August 2008
- HMIC (Ovid): 2000 - 2008
- MEDLINE (Ovid): 2000 – 2008
- NHSEED: 2000 - 2008

- Social Science Citation Index (Web of Science): 2000 – August 2008
- Social Care Online: 2000 - 2008

These searches were restricted to studies that explicitly named a service being evaluated as intermediate care or as a service widely recognised to be a form of intermediate care. The search was limited to literature published in English on evaluations of services conducted in the UK and did not consider unpublished work. Publications were identified by their abstracts or, where these were unavailable, their titles and authors. Each publication was evaluated to determine its relevance and apparent importance to the review after the guidance of Roberts et al (2002). Where possible, copies were retrieved for more detailed evaluation. These articles, books or monographs were read and appraised and where possible, coded.

2.5.4.2 How papers were selected for review

Stage I - Initial Scoping Search

Following an initial scoping exercise, searches identified were loaded into Reference Manager. References that came up from the identified papers were then combed for additional relevant publications using the Social Sciences Citation index. Most of the searches in the computer databases were re-run in a bid to identify newly published papers. Searching electronically also helped to identify some forthcoming papers. The following criteria were used at this stage:

- Study design: any evaluation, analysis, assessment, investigation, survey, review or research.

- Population: people 65 years and older.
- Setting: the UK.
- Intervention: a service that has been explicitly named as intermediate care or as a service widely recognised to be a form of intermediate care.
- Outcomes: quantitative outcomes including costs and outcomes of utilisation of intermediate care services (quality of life, functional status, discharge destination etc).
- Reporting: all evaluative studies, excluding duplicates.

Three more stages then followed.

Stage II - Inclusion and exclusion criteria

Only abstracts were read at this stage. A predetermined and explicit criteria was used to make the inclusion and exclusion decisions. Studies were only included if they were quantitative evaluations of intermediate care services. Studies that were clinical trials, or parts thereof, were also included.

Stage III – Categorisation of studies

Full papers of each of the studies deemed to be relevant after stage II were then read. Each study was categorised on the basis of its title, Medical Subject Heading (MeSH)

and abstract, when available. The following criteria were then used to determine the relevance of each study to the literature review.

- A. Primary research involves a clinical trial.
- B. Primary research involves descriptive study.
- C. Primary research involves non-experimental analytic study.
- D. Systematic review of studies that fall in categories A, B or C.
- E. Study does not have any relevance to the evaluation of intermediate care.

Using stage III criteria, studies that were coded as either A, B, C or D were considered to be relevant to the review. Those classed as E were not deemed appropriate and were therefore not considered further. At this stage, articles were not graded in terms of research quality, or removed from the review for reasons of poor quality, although many had problems of design and reporting. This was because even where the goal and methodology were unsuitable to answer the question posed by this evaluation, a study had often generated useful information with which to address one or more of the evaluation issues of interest to the review.

Stage IV - data extraction

The full papers or studies chosen after stage III were read and information extracted into a table format. The following information was recorded about each relevant

study: setting of study, study design, type of data collected, what was evaluated and key findings (Table 2.1).

2.5.4.3 Results

Stages I, II, III and IV

A total of 89 papers were originally selected from the initial scoping exercise (Stage I). At stage II, these studies were reduced to 34 papers, and to 31 upon reading the abstracts. During stage III, full papers of each of the 31 studies were read and these studies were then categorised into four groups (A, B, C and D) described on page 31. Data was then extracted from these 31 studies. The evaluations have all been quantitative in nature. Table 2.1 gives a summary of studies reporting such evaluations.

Description of Studies

All studies identified were of evaluations done in a UK setting. Eighteen of the studies were trial-based (controlled or uncontrolled). All the trials had either the control or the intervention group based in a hospital or GP setting with the exception of three (Fleming et al. 2004; Gunnell et al. 2000; Wade et al. 2003). One of the studies was descriptive, analysing data from a home-based intermediate care service (Beech et al. 2004). Another 11 studies were non-experimental analytic studies. The non-experimental analytic studies reported evaluations of larger samples drawn from intermediate care services, agencies, schemes or projects. There was one literature review of three trials reported (Wilson-Barnett et al. 2001).

Outcomes assessed

Various types of variables were analysed in the trials including outcome measures (quality of life, functional status and mental health measures), cost data, length of stay, discharge destinations, mortality, staff ratios, readmission rates, measures of patient satisfaction, risk of fall, demographic and medical characteristics, staff and patient views, user experiences and performance indicators.

2.5.4.4 Key Findings

The benchmark for gauging or assessing the performance of intermediate care services differed according to the type of study that was conducted. In the 21 clinical trials considered (three of these were part of the review conducted by Wilson-Barnett et al. 2001), the comparator was usual, conventional, general or standard care. In one study, the same intermediate care service was evaluated at two time periods (Wilson et al. 2003). There was no comparator for the descriptive study as it merely reported various characteristics of a multidisciplinary rapid response team. For the majority of non-experimental analytic studies, there was no explicit comparator even though the results were usually analysed in implicit comparison to alternatives of intermediate care. One of these analytic studies compared different forms of intermediate care against each other (Kaambwa et al. 2008). In many cases, multiple outcomes were analysed for a single study. The key findings from the evaluations are reported below according to the study design and the type of outcomes considered.

Clinical trials

Three of these trials (Miller et al. 2005; O'Reilly et al. 2006; O'Reilly et al. 2008) conducted cost-effectiveness analyses using quality-adjusted life years (QALYs) and

costs as outcomes. In Miller et al (2005), an early discharge and rehabilitation service in Nottingham was found to be more cost effective than usual care while the other two concluded that cost-effectiveness of post-acute care provided in community hospitals was similar to that provided in general hospital care. Other trials focussed on effectiveness analyses and compared various outcomes. Wilson-Barnett et al (2001) reviewed the results of three trials and found that two of these (Griffiths et al. 2000; Griffiths & Wilson-Barnett 1998) reported that nursing-led inpatient units (NLIUs) had slightly lower daily costs than usual care. A later trial (Griffiths et al. 2001) also concluded that a NLIU was associated with lower daily costs compared to usual care. However, three trials found the opposite to be true with NLIUs having higher costs than usual or standard care (Harris et al. 2005; Richardson et al. 2001; Walsh et al. 2005).

Other trials focussed on the Barthel index or other measures of independence as outcome measures. Compared to their comparators, five of these trials found that intermediate care services were associated with better Barthel scores (Harris et al. 2007), physical outcomes (Boston et al. 2001) or other measures of independence (Green et al. 2005; Griffiths & Evans 1995; Young et al. 2007). Three trials did not find any difference between intermediate care services and their comparators in Barthel scores (Griffiths et al. 2000), physical outcomes (Boston et al. 2001) or institutionalisation (Fleming et al. 2004). Another study found that intermediate care was associated with poorer health outcomes as measured by the SF-36 (Wade et al. 2003).

Of those that compared length of stay, two found intermediate care services to be associated with shorter length of stay (Cotton et al. 2000; Griffiths & Evans 1995), while four reported longer lengths of stay for intermediate care services (Griffiths et al. 2001; Harris et al. 2007; Griffiths et al. 2000 and Griffiths & Wilson-Barnett 1998). Only one trial reported similar lengths of stay for intermediate care and usual care (Richardson et al. 2001).

In terms of discharge destinations, two studies were associated with favourable destinations compared to usual care (Harris et al. 2007 and Griffiths & Evans 1995) while no evidence of better destinations was found in another (Griffiths et al. 2001). Delivery of intermediate care services did not result in reduced need for readmission compared to usual hospital care (Cotton et al. 2000) while Wilson et al. (2003) found no difference between the trial and post-trial outcomes of a hospital at home intermediate care scheme. The only trial that focused on carer strain did not find any evidence of a difference between intermediate care and usual care (Gunnel et al. 2000).

Descriptive study

A survey of patients' and carers' satisfaction with intermediate care found that the majority of them (patients and carers) had positive views about these services (Beech et al. 2004). The majority of users were older people with mean age 75.9 years and 57% had both medical and social care needs.

Non-experimental analytical studies

In terms of costs, Mayhew and Lawrence (2006) concluded that proper implementation of guidelines would result in cost savings. In Kaambwa et al (2008), intermediate care services that performed an admission avoidance service had lower costs than those that performed a supported discharge function. The respiratory intermediate care team in Ward et al (2005) was linked with cost savings.

Peet et al (2002) concluded that intermediate care was associated with better functional (Barthel and extended activities of daily living) and health (EQ-5D) outcomes while admission avoidance services had greater EQ-5D and Barthel gains than supported discharge services in Kaambwa et al. (2008)

Intermediate care was not effective in reducing emergency admissions for older people in one study (Walker & Jamrozik 2005) while it led to an increase in emergency admissions in another (Walker et al. 2005). Some of the admissions or referrals to intermediate care had been inappropriate (Kaambwa et al. 2008; Peet et al. 2002) while there was evidence that certain groups of patients had been denied access e.g. those with cognitive impairment (Carpenter et al. 2003). This all pointed to the need to establish well-defined admission criteria.

The diversity with which intermediate care is applied in various parts of the UK was shown (Institute of Health Sciences and Public Health Research 2005). There was also huge pressure placed on intermediate care services e.g. on rehabilitation wards (Carpenter et al. 2003).

2.5.4.5 Messages from literature review

Results from studies such as Beech et al (2004) and Institute of Health Sciences and Public Health Research (2005) epitomise the performance of most intermediate care services in the UK and bring home the message that intermediate care evaluation results were both positive and negative. On one hand, intermediate care services were seen as beneficial as they complemented the mainstream services but on the other hand, they were a barrier to seamless services because of several reasons including their bureaucracy (Cornes & Clough 2001). Other evaluations however revealed that there was no difference between intermediate care and alternative services. These results have led some to conclude that there is no 'scientific' evidence that exists about the benefits of intermediate care (Melis et al. 2004). Further, information on costs and cost effectiveness continues to be inadequate (Beech 2005). Inadequacies in the performance of intermediate care services have led to conclusions that a lot more needs to be done in terms of increasing physical and staff capacity (Peet et al. 2002; Carpenter et al. 2003), increasing influence over choice of care and the quality of information about care (Beech et al. 2004) and increasing the proportion of minority ethnic groups being cared for (Peet et al. 2002).

Beech et al (2004) submit that there is need for more research to be conducted so as to ascertain whether or not intermediate care services can be seen as effective, suitable and efficient alternatives to acute care. The recommendation for more evaluation was also made by Fleming et al (2001). A recommended methodological approach to deal with the difficulty of using scientific methods when evaluating intermediate care services is to have five phases of evaluations: the first two phases would be descriptive and generate information for subsequent phases, phases three and four

would focus on formal evaluations using scientific methods while the last phase would monitor the ongoing effects of successful interventions (Beech 2005).

2.5.5 Key messages for health services researchers from this chapter

The population of older people in the UK, like in the rest of the world, has been increasing in both absolute and relative terms. In response to the unique needs of older people, various models of care have been adopted in the UK and intermediate care is one of them. There is considerable debate about the most appropriate definition of intermediate care and various forms of intermediate care exist. The literature review of studies that have evaluated intermediate care services in the UK revealed mixed results about the effectiveness and cost-effectiveness of these services. There is thus need for more evidence to be collected so as to assess the performance of intermediate care services.

2.6 Conclusion

This chapter has traced the global and UK demographic trends for older people. What is obvious from this exercise is that the proportion of older people has been increasing globally and will continue to do so in the future. An older population brings with it distinctive challenges due to its unique needs and requirements such as health, mental and usual activities. There have been several government initiatives and policies in the UK to meet the needs of older people. One initiative has been encapsulated in the eight standards of the NSF (Department of Health 2001b). Intermediate care is one of these standards. This chapter has reviewed various definitions of intermediate care and presented different models of intermediate care. Notable milestones in the development of intermediate care in the UK have also been charted and the chapter

has ended by conducting a literature review of notable studies where services that can be explicitly identified as intermediate care have been evaluated. The evidence from these evaluations is that intermediate care has been associated with both good and bad outcomes when compared to other forms of care. The need for more evaluations to be conducted to create a bigger information base can never be overemphasised.

The next chapter expands this literature review by considering evaluations of all services for older people in the UK (intermediate care included) with a view to identifying key statistical issues that are commonly associated with such evaluations and that often lead to bias in statistical analysis. As such, only evaluations with a quantitative element are considered.

Table 2.1: Evaluations of Intermediate Care

Reference	Setting	Study Design	Type of data collected	What was evaluated	Key Findings
Trials					
Boston et al (2001)	An Inner London Health Authority	Prospective non-randomised comparative study	Demographic and medical data; cognitive function (Abbreviated Mental Test); mental state (Philadelphia morale scale); SF-12; quality and quantity of support	Comparison of elderly patients admitted to an inner city GP unit with comparable patients in conventional care	Physical outcomes in GP units and conventional settings were similar. GP units were associated with short-term improvement in mental functioning and better quality of care.
Cotton et al (2000)	Glasgow Royal Infirmary, UK	Randomised controlled trial	Readmission; additional hospital days, deaths within 60 days of initial admission	Comparison of early discharge with home treatment of patients supported by respiratory nurses with usual hospital care	Early discharge policy was associated with reduced inpatient stay. There was no difference in the subsequent need for readmission between the two groups.
Davies et al (2000)	University teaching hospital, Liverpool, UK	Prospective randomized controlled trial	Readmission rates; changes in forced expiratory volume in one second (FEV ₁); Mortality	Comparison of 'hospital at home' and hospital care as an inpatient for patients with an exacerbation of chronic obstructive pulmonary disease	Hospital at home was found to be a practical alternative to usual (hospital) care for certain patients.
Fleming et al (2004)	Rehabilitation service based in social Services old people's homes in Nottingham, UK	Randomised Controlled Trial	Institutionalisation rates; Barthel Index; Nottingham Extended ADL scores	Evaluation of the effect of a care home rehabilitation service on institutionalisation, health outcomes and service use	There was no reduction on institutionalisation due to the service compared to usual care.
Green et al (2005)	Community hospital and district general hospital in Bradford, UK	Randomised controlled trial	Nottingham extended activities of daily living scale; general health questionnaire 28; Barthel Index; Nottingham health profile.	The effect that a locality based community hospital has on independence in older people needing rehabilitation compared to usual care in a hospital ward	There greater association between independence and care in a locality community hospital compared to care in a district general hospital.
Griffiths et al (2000)	Inner London Hospital Trust	Randomised Controlled Trial	Barthel Index; Length of stay; discharge destinations; mortality; nurse-patient ratios	Assessment of the potential for a nursing-led-inpatient unit (NLIU) – comparison between plan to transfer to NLIU and plan to remain under usual care	No significant difference in Barthel index between NLIU and usual care. NLIU associated with longer length of stay and lower nurse-patient ratios.
Griffiths et al (2001)	Hospital wards in an acute inner London National Health Service Trust	Randomised Controlled Trial	Length of stay; costs; discharge destination; Barthel Index	Determination of the outcome and cost of transferring patients to a nursing-led inpatient unit for intermediate care compared to usual hospital care.	NLIU associated with longer length of stay, lower daily cost and higher mean costs, NLIU had no significant effect on discharge destination.

Table 2.1: Evaluations of Intermediate Care

Reference	Setting	Study Design	Type of data collected	What was evaluated	Key Findings
Trials					
Gunnel et al (2000)	Bristol, UK	Randomised controlled trial	Modified 12-item Carer Strain Index; COOP-WONCA charts; EuroQol EQ5D	Assessing the impact of early discharge hospital-at-home scheme (compared to hospital) on carer's strain and quality of life	There was no evidence that self-reported of increased strain on carers of patients discharged from hospital.
Harris et al (2005)	UK nursing-led Inpatient Unit	Randomised Clinical Trial	Barthel Index, Cost per Day, length of stay	Costs from the UK National Health Service perspective, of transfer to a nursing-led inpatient unit for intermediate care	Nursing-led inpatient units have higher costs of hospital stay (despite having a lower cost/day) than acute hospital wards because of longer length of stay in the former.
Harris et al (2007)	Three purposefully replicated, pragmatic randomized controlled trials	Pragmatic randomized controlled trials	Barthel Index; length of stay; discharge destination; mortality; General Health Questionnaire; Nottingham Health Profile Distress Index (NHPDI) ; incidence of complications and readmission	Determination of whether transfer to a nursing-led inpatient unit (NLIU) prior to discharge from hospital can improve clinical outcome and reduce length of stay and readmission rate for medically stable post-acute patients assessed as requiring inpatient care.	Post-acute patients with complex health and social needs transferred to a NLIU can have better outcomes of care (Barthel; General health questionnaire, NHPDI; pressure ulcers) but longer length of stay.
Miller et al (2005)	Acute and rehabilitation wards in NHS hospitals in Nottingham	Randomised Clinical Trial	Costs, EuroQol EQ-5D, Quality adjusted life years (QALYs)	The cost-effectiveness of an early discharge and rehabilitation service (EDRS) compared to usual care.	The Nottingham EDRS was more cost effective than usual care
O'Reilly et al (2006)	Community hospital and district general hospital in Yorkshire, UK	Randomised Clinical Trial	Costs, EuroQol EQ-5D, Quality adjusted life years (QALYs)	Cost effectiveness of post-acute care for older people in a community hospital compared with care in a district general hospital department for older people with acute medical conditions	There is no difference in cost-effectiveness between the two groups.
O'Reilly et al (2008)	Seven community hospitals and five general hospitals at five centres in the midlands and north of England.	Randomised Controlled trial	EuroQol EQ-5D and health and social service costs	A comparison of the cost effectiveness of post-acute care for older people provided in community hospitals with that provided in general hospital care.	The cost-effectiveness of post-acute provided in community services was similar to that provided in general hospital care.

Table 2.1: Evaluations of Intermediate Care

Reference	Setting	Study Design	Type of data collected	What was evaluated	Key Findings
Trials					
Richardson et al (2001)	Nursing-led intermediate care unit in an inner London teaching hospital	Randomised Controlled Trial	Length of stay; costs	Assessment of the cost and impact on outcomes of introducing a nursing-led ward program	No significant difference in outcomes of NLIU and standard care. Costs for NLIUs are higher because of longer length of stay.
Shaw et al (2003)	Two A&E departments, Newcastle upon Tyne, UK	Randomised controlled trial	Number of fallers; number of falls; time to first fall; injury rates; Mortality; fall related hospital admissions fall related attendances at A&E department	The effectiveness of multifactorial intervention after a fall in older patients with cognitive impairment and dementia attending the accident and emergency (A&E) department	Multifactorial intervention was not effective in preventing falls in older people.
Wade et al (2003)	Oxford, UK	Randomised controlled (crossover) trial	EuroQoL EQ-5D; SF-36; Parkinson's disease disability questionnaire; Parkinson's disease questionnaire ; hospital anxiety and depression scale; limited stand-walk-sit test; carer strain index	Whether or not a programme of multidisciplinary rehabilitation and group support is associated with sustained benefit for people with Parkinson's disease or their carers	A short spell in multidisciplinary rehabilitation may improve mobility (stand-walk-sit test) but this service had worse general and mental health (SF-36).
Walsh et al (2005)	Nurse-led unit and acute general medical wards in a large urban UK teaching hospital	Pragmatic Randomised Clinical Trial	Costs	Cost minimisation study of nurse-led intermediate care compared with standard hospital care for post-acute medical patients	Nurse-led intermediate care more expensive than standard hospital care.
Young et al (2007)	Community hospital and district general hospital in Bradford, UK	Randomised controlled trial	Time to transfer; Nottingham extended activities of daily living scale	Investigation of the effect of delayed community hospital transfer on outcome	Shorter time to community hospital transfer was associated with improved independence.

Table 2.1: Evaluations of Intermediate Care

Reference	Setting	Study Design	Type of data collected	What was evaluated	Key Findings
Descriptive Studies					
Beech et al (2004)	A multidisciplinary Rapid Response Team (RRT) in Herefordshire Primary Care Trust	Descriptive study	Views on the new service, number of service users, medical and social needs	Evaluation of the RRT	On the whole, patients and carers were positive about RRT but there also disquiet about the lack of influence over choice of care and the quality of information about care.

Table 2.1: Evaluations of Intermediate Care

Reference	Setting	Study Design	Type of data collected	What was evaluated	Key Findings
Non-experimental analytic studies					
Brooks et al (2003)	A Rapid Assessment Support Service (RASS) in the UK.	Non-experimental analytic study	Discharge destinations; readmission to hospital; patient needs	Exploring if Intermediate Care could reduce number of unnecessary emergency hospital admissions	The RASS appeared to have prevented unnecessary admissions to hospital for older people. Only 5% were readmitted to hospital.
Carpenter et al (2003)	Shepway District of Kent, UK	Non-experimental analytic study	Staff and bed numbers; admission criteria; resource use; Patients' medical, physical, psychological and social characteristics.	An evaluation of intermediate care services (ICSs) for older people by examining the relationship between different intermediate care services and the use of hospital beds and intermediate care resources.	Patients admitted to ICSs had different characteristics and some of these did not seem to meet the criteria. Rehabilitation wards were under pressure from high patient demand while the cognitively-impaired were denied access to ICSs. There was need to increase the capacity in the community-based ICSs as well as have a well-defined admission criteria for ICSs.
Institute of Health Sciences and Public Health Research (2005)	Primary Health Care trusts from North England	Non-experimental analytic study (Comparative case study)	Costs; performance indicators; Service delivery, culture and behaviour data; User experiences and outcomes	Examining the structure, process, outcomes and cost effectiveness of intermediate care for older people with a focus on the impact at three levels: service system, service components and individual patient/user and caregiver	Intermediate care is conceived and implemented in diverse ways. High level intermediate care performance indicators did not offer either positive or negative evidence. Intermediate care has however been associated with substantive changes in the structure of service delivery and in the culture and behaviour of commissioners and providers.
Kaambwa et al (2008)	Five anonymous case studies in the UK	Non-experimental analytic study	Costs, Barthel index and EQ-5D	Comparison of costs associated with different intermediate care functions (supported discharge, admission avoidance) and settings (residential and non-residential)	Almost 50% of patients were inappropriately admitted to Intermediate care. In comparison to supported discharge, admission avoidance services were associated with both lower costs and greater health and functional gains.
Mayhew and Lawrence (2006)	London Borough of Brent with a population of 260,000	Non-experimental analytic Study (Workshops)	Costs of IC packages & acute hospital admission; physical resources	An estimation of the changes on provision and costs of intermediate care as a result of reductions in acute hospital care through prevented admission and early discharge	Sound implementation of in intermediate care would result in cost saving to the health economy.

Table 2.1: Evaluations of Intermediate Care

Reference	Setting	Study Design	Type of data collected	What was evaluated	Key Findings
Non-experimental analytic studies					
Peet et al (2002)	Eight admission avoidance, early discharge and community reablement schemes in Leicester city; Leicestershire and Rutland	Non-experimental analytic study (simple comparative study)	Patient's length of stay; Barthel ; source and reasons for referral; extended activities of daily living; discharge destination; EuroQol scores	Understanding the processes, outcomes and costs of a number of alternatives to the eight schemes and identifying elements of best practice of intermediate care in Leicester and Rutland	The majority of intermediate care service users were discharged to their own homes and also had major improvements in their functional and health status. Overall outcomes for all service users were positive. Some problems identified were staff recruitment, inappropriate referrals, few users from ethnic minorities and problems with the physical environment.
Walker and Jamrozik (2005)	An intermediate care project in a West London Primary Care Trust	Non-experimental analytic (cohort) study	Levels of coverage	Evaluation of the effectiveness of screening for risk emergencies in the Keep Well At Home (KWAH)	The KWAH project has not been effective in reducing emergency admissions for the elderly.
Walker et al (2005)	An intermediate care project comprising 20 of the 38 practices in a West London Primary Care Trust (Hammersmith and Fulham Primary Care Trust)	Non-experimental analytic (cohort) study	Patterns of emergency care	The efficacy of the Keep Well At Home (KWAH) project by determining the use of Accident & Emergency (A&E) services by those who had been screened by the project.	There was a 51% increase in the crude rate of emergency admissions in the first year after screening compared to 12 months before assessment which was not expected.
Ward et al (2005)	A respiratory intermediate care team (RICT) in Oxford City Primary Care Trust	Non-experimental analytic study (Audit & patient satisfaction postal questionnaire)	Costs; Reason for referrals, source of referral; patient satisfaction views	The effectiveness of the work of RICT	The RICT is effective in saving hospital bed days through prevention of admission and early discharge of chronic obstructive pulmonary disease (COPD) patients. Hospital at home was found to be more favourable. The RICT led to cost savings on hospital admissions.

Table 2.1: Evaluations of Intermediate Care

Reference	Setting	Study Design	Type of data collected	What was evaluated	Key Findings
Non-experimental analytic studies					
Wilson et al (2003)	Admission-avoidance hospital-at-home scheme in Leicester, UK.	Non-experimental analytic study	Baseline characteristics; survival; Barthel index, Sickness Impact Profile 68, Philadelphia Geriatric Morale Scale; Length of stay and visits from a general practitioner	Comparison of the performance of an admission-avoidance hospital-at-home scheme one year after the end of a randomised trial with performance during the trial	There were no significant difference in the results of the trial and post-trial periods except for higher volume of work and as shorter length of stay for the latter.
Young et al (2003)	An elderly care department in North Bradford Primary Care Trust, UK	Non-experimental analytic (Prospective) study	Post-acute needs (Rehabilitation, new care home, palliative care, respite/convalence care)	Estimating the need for post-acute intermediate care in an elderly department for older people	Post-acute intermediate care should have capacity to cater for the needs of up to a quarter of acute admissions to a district general hospital elderly care department.

Table 2.1: Evaluations of Intermediate Care

Reference	Setting	Study Design	Type of data collected	What was evaluated	Key Findings
Other Studies					
Wilson-Barnett et al (2001)	Nurse-led inpatient units (NLIUs) in south and north London, UK	Literature Review	Length of stay; number of complications; independence; number of discharges; costs;	A review of three studies that evaluated NLIUs in the UK	The first study (Griffiths and Evans, 1995) showed that NLIUs has shorter stay, fewer complications; higher levels of independence and fewer discharged to nursing homes The last two studies (Griffiths et al. 2000; Griffith and Wilson-Barnet 1998) showed that compared to usual care wards, NLIUs were associated with a longer stay and slightly lower daily costs.

CHAPTER THREE - LITERATURE REVIEW OF COMMON STATISTICAL PROBLEMS IN EVALUATIONS OF SERVICES FOR OLDER PEOPLE

3.1 Introduction

The aim of this chapter was to review studies that have evaluated services for older people with a view to identifying key statistical issues that are commonly associated with such evaluations. Identified statistical problems would form a framework of common statistical issues one has to contend with when evaluating services for this age group. The key messages for health services researchers in this and other related areas of study are also given. This work, in part, builds on that by Parker et al (2000) which was a systematic review that focussed on evaluative research literature on the costs, quality and effectiveness of different locations of care for older patients. The focus of the Parker et al. review was on policy issues whereas the key aim of this chapter is identifying methodological issues. A summary of the Parker et al. review is presented in section 3.3 and similarities and key differences between the Parker et al review and the review conducted in this chapter are outlined.

3.2 Definitions

This study focuses on ‘service evaluations’ and therefore an understanding of what is meant by the phrase is important.

In the Collins English Dictionary (2000), service is ‘...an act of help or assistance or an organized system of labour and material aids used to supply the needs of the public or to supply with assistance’.

According to Rossi and Freeman (1993, p.5), evaluation is "...the systematic application of social research procedures for assessing the conceptualization, design, implementation, and utility of ... programs." Rossi and Freeman (1993) posit that evaluations can be further conceptualised as:

- (i) Process evaluations which describe and assess service activities while that particular service or programme is running.
- (ii) Outcome evaluations which assess the immediate or direct achievements and effects of a service.
- (iii) Impact evaluations which consider long-term as well as unintended service effects, thereby looking beyond the immediate results of policies, instructions or services.

Service evaluations for older people can therefore be operationally defined as assessments of organised activities that supply labour and material aids for the assistance of older people.

3.3 Summary of Parker et al (2000)

This systematic review assessed studies that have evaluated services catering for older patients. The outcomes considered were costs, the quality and effectiveness of different locations of care for older patients. The locations investigated were for acute care, post-acute care, sub-acute care and rehabilitation care for people aged 65 years and over. A number of databases covering the period 1988 to 1999 were searched.

Though many key outcomes were covered in the full review (Parker et al. 1999), only four were considered for the subsequent analysis whose results are reported in Parker et al (2000): mortality, destination outcome, re-admission and costs to the health service. The analysis of each of the four outcomes was conducted for each of the nine locations of care that fell under three broad categories:

- services that adjusted skill mix (admission avoidance services, nurse-led beds and supported early discharge services),
- hospital settings that had increased condition-specific expertise (stroke units, hip units and geriatric assessment units or acute care for elders units and
- services that offered rehabilitation (inpatient rehabilitation, community-based rehabilitation and day hospitals).

Firm conclusions could not be arrived at except for stroke units, early discharge services and geriatric assessment units. They further pointed out that evidence on cost-effectiveness, let alone the effectiveness, of intermediate care is lacking in these studies. They also raised questions about the appropriateness of systematic review techniques in service delivery and organisation.

3.3.1 Similarities with the review in this chapter:

- The focus of the review was on literature on service evaluations.
- The sample was of older people aged 65 years and over.
- It also considered some quantitative outcomes i.e. costs and effectiveness.
- Some of the databases searched are also used in this chapter.
- Search strategy involved an iterative process.

3.3.2 Differences from the review in this chapter:

- Parker et al (2000) did not have a restriction in the searching by the country of origin while this chapter looked at studies in the UK only.
- Only studies with a randomised or pseudo-randomised trial design were included in Parker et al (2000) while this chapter considers all ‘quantitative’ studies.
- The search period in Parker et al (2000) was from 1988 while this chapter begins from 2000.
- The focus of Parker et al (2000) was on policy issues while this chapter focuses additionally on methodological aspects of the studies.

3.4 Methods for the Literature Review

3.4.1 Sample

Services that cater for the treatment and rehabilitative needs of older people may be institutional based (such as geriatric wards or residential homes) or community-based (such as rapid response teams) (Department of Health 2001a). Care for older people is conducted in so many places and in different settings and therefore doing this literature review presents some challenges. It is therefore of great importance that a clear and consistent approach is used to describe such services. In chapter two, only intermediate care services were considered. Here, the scope of services was expanded to include all services that cater for older people aged 65 and over in the UK. A more detailed description of the criteria is given under the ‘How papers were selected’ section below.

3.4.2 Search Strategy

The strategy used in chapter two (section 2.5.4.1) was expanded and used here. The same range of online computer databases was searched to identify relevant literature.

The search of databases carried out involved combining the terms shown in A1 and A2 of the appendix and searching for material from 2000 onwards. This starting point was chosen to ensure that the focus was on UK studies which had published since the “Best Place of Care” review was conducted in 1999 (Parker et al. 2000). These searches encompassed literature relating to several conditions associated with older people such as Alzheimer’s disease, depression, dementia, strokes and falls.

3.4.3 How papers were selected for review

Stage I - Initial Scoping Search

Stage I was the same as that used in chapter two save for one difference. This was that the intervention considered was ‘all services’ that are used exclusively by older people. Intermediate care was included among these services.

Stage II - Inclusion and exclusion criteria

This was the same as that in chapter two except that only IC studies were considered.

Stage III – Initial categorisation of studies

Upon reading the abstracts, the following criteria were then used to determine the relevance of each study to the literature review:

- A. Primary research is on costs, utilisation and outcome data of services for older people and formal evaluation or hypothesis testing was conducted.

- B. Focus is on services for older people (but not a formal evaluation or hypothesis testing). Useful primary and secondary cost, utilisation and outcome data included.
- C. Systematic review of studies that fall in categories A or B.
- D. Contains useful information but does not obviously fall into categories A or B.
- E. Study does not have any relevance to the evaluation of services for older people.

Using stage III criteria, studies that were coded as either A, B or C were considered to be relevant to the review. Some studies coded as D were also included. Those classed as E were not deemed appropriate and were therefore not considered further. Examples of studies that were deemed to be inappropriate and therefore excluded, together with the reasons for this exclusion, are given in Table 3.1 below.

Stage IV – further categorisation of studies

An intermediate stage was introduced in this strategy. After reading the full papers, all studies in categories A, B, C and (or) D were further classified by type of study into the following categories:

1. Economic evaluation (cost-minimisation analysis, cost-effectiveness analysis, cost-utility analysis or cost-benefit analysis) in a clinical trial.

2. Other evaluation or analysis in a clinical trial.
3. Other cost study.
4. Quantitative evaluation or analysis of both costs and other outcomes for older people.
5. Quantitative evaluation or analysis of other outcomes of services for older people (other than costs).
6. Description of methods used in aspects of evaluation of services for older people.
7. Not relevant on evaluation of full paper.

All studies classified as A(1), A(2), A(3), A(4), A(5), B(1), B(2), B(3), B(4), B(5), C(1), C(2), C(3), C(4), C(5), D(1), D(2), D(3), D(4), or D(5) were included in the next stage. Some studies classified as A(6), B(6), C(6) and D(6) were also included.

Stage V - data extraction

The same procedure conducted in Stage IV of chapter two was repeated here. The following information was recorded about each relevant study: primary focus of research, type of population, sample size, study design (including details of key analyses conducted), data sources, key outcomes and key findings (Table 3.2). Table 3.3 then summarises the statistical issues or problems identified in the studies.

3.5 Results

Stages I, II, III, IV and V

A total of 3155 papers were originally selected from the initial scoping exercise (Stage I). At stage II, these studies were reduced to 553 papers, and to 72 at stage III upon reading the abstracts. During stage IV, full papers of each of the 72 studies were read and these studies were then categorised into further groups. 16 studies were also then excluded at this stage. Data was then extracted from 56 studies.

3.5.1 Description of studies from whom data was extracted

All studies reported in Table 3.2 had populations of older people i.e. aged 65 years or over. Some studies that have been widely referenced elsewhere were however omitted from this table because they had lower age floor limits: e.g. 18 years (Griffiths et al. 2005;Shepperd & Iliffe 2005), 45 years (Bajekal et al. 2004), 50 years (Roe et al. 2003), 52 years (Ebrahim et al. 2004), 55 years (Burholt 2004;Ellis et al. 2006;Roderick et al. 2001;Spiers et al. 2005), 60 years (Evandrou 2000;Scuffham, Chaplin, & Legood 2003;Smith et al. 2004) and 63 years (Young et al. 2005a;Young et al. 2005b). Some studies clearly dealt with populations of older people but did not specify the age range (e.g. Walsh et al. 2005, Oliver et al. 2004 and Forder & Netten 2006). The sample sizes varied from a minimum of 57 to a maximum of 514,420 observations. All data was from UK sources such as intermediate care schemes, rehabilitation settings, geriatric hospital wards, hospitals, general practices, primary care trusts, and various controlled trials and non-experimental analytic studies. Diverse findings are reported in table 3.2. Further, all the studies reported can be divided up into three of the five broad categories presented in ‘Stage III – Initial categorisation of studies’ i.e. (A, B or C).

Category A studies

Fifty studies fell under category A (studies whose primary research was on costs, utilisation and outcome data of services for older people and formal evaluation or hypothesis testing was conducted). The majority of these studies were controlled trials (21 studies) with five of them focussing on economic evaluation. The rest of the controlled trials analysed costs and outcomes outside an economic evaluation framework with three using regression analysis while the remainder employed other statistical approaches in their analyses. The other studies in this category were classified as A3, A4, A5 or A6. The bulk of these four categories of studies analysed observational data (e.g. longitudinal cohort datasets) while the rest examined routinely collected datasets, secondary data (e.g. data from previous surveys) or data from face to face interviews. By far the most used form of analysis in these four categories was simple statistical analysis followed by use of a regression model. Other approaches used included maximum likelihood (Matthews & Brayne 2005) and a computer simulation model (Carpenter et al. 2003).

Some of the key outcomes analysed in category ‘A’ studies included costs, length of stay, mortality, morbidity, referrals, general practitioner (GP) and hospital utilisation patterns, number of falls, discharge destinations, institutionalisation rates, incidence of complications, market indicators, nurse-patient ratios, psychological and social characteristics, patient demographic characteristics and socio-economic indicators. In addition, various measures of functional independence were reported (e.g. Barthel index, Rivermead mobility index, Frenchay activities index and the Nottingham extended activity of daily living score) as well as measures of mental health (abbreviated mental test – AMT, mini-mental state examination, Eysenk personality inventory and the depression score). Other outcomes were quality of life indicators

(EQ-5D, quality adjusted life years, sickness impact profile, Philadelphia geriatric morale scale, Short Form 36, Western Ontario and McMaster Osteoarthritis Index – WOMAC, general health questionnaire, Nottingham Health Profile Distress Index and Satisfaction with life scale).

Category B studies

Only one study (O'Shea 2004) was classified under category 'B' (studies that focussed on services for older people that did not involve any formal evaluation or hypothesis testing but include useful primary and secondary cost, utilisation and outcome data). It was a descriptive study that used statistical tests in its analysis. Outcomes analysed included costs and indicators of best practice.

Category C studies

Category 'C' was made up of five studies that conducted a systematic review of other evaluations. Three of the five studies (Parker et al. 2000; Elkan et al. 2001 and Wilson-Barnett et al. 2001) were reviews of controlled trials. Elkan et al (2001) performed a meta-analysis; Parker et al (2000), among other analyses, used the Mantel-Haenszel method while the others performed other statistical analyses. Data sources included Medline, Cinahl, Web of Science, Cochrane Library, Science Citation Index, Psychlit, reference lists of papers, abstracts from scientific meetings, Effective practice and organisation of care specialist register, HMIC, British nursing index as well as hand searches. The most common outcomes considered were mortality, discharge destination outcomes, re-admission rates, costs, living arrangements, admission rates, functional and health measures.

3.5.2 Statistical problems identified in studies

The exercise revealed that a lot of studies have statistical challenges that a researcher needs to account for when conducting an analysis. A list of the common statistical problems from the studies is presented in table 3.3. These statistical problems are presented below in order of their prevalence in the studies which were part of the literature review.

3.5.2.1 Missing data

By far the most common problem was that of missing data and/or attrition reported by 29 of the studies. Missing data is a very prevalent phenomenon in studies especially those that involve older people (Tomassini 2006). Missing data leads to biased results (Schafer 1997) and loss of power (Roderick et al. 2001). Many reasons were advanced as leading to missing data including non-response (Breeze et al. 2003; Harris et al. 2005), participants' health decline (Beaumont & Kenealy 2004), migration (Seshamani & Gray 2004a; Seshamani & Gray, 2004b), death¹ (Beaumont & Kenealy 2004), plain refusal to take part in a survey (Mann et al. 2006) or attrition (Matthews et al. 2004). Attrition refers to the loss of relevant individuals after the initial definition of the population that is to be included in a study (Matthews et al. 2004). This results in systematic differences between the comparison groups due to this loss of participants. In investigating bias that results from attrition, Matthews et al (2004) named non-response as the first stage of attrition followed by dropout. Four types of dropout were discussed: death of participants, failure to contact participants, inability of participants to respond and, lastly, their refusal to respond.

¹ In health economics studies, such as in the analysis of quality of life data, there is a qualitative difference in the way 'death' and other sources of missingness are treated. When using the EuroQoL EQ-5D for instance, dead people will be assigned a value '0' and will therefore not be treated as 'missing' (Diehr & Johnson 2005).

Suspected initial non-response in a sample of 2,500 individuals in Mathew et al (2004) was tested by comparing individuals whose data were not collected at baseline to that of individuals for whom data were collected in the two-year follow-up interviews. About 2.9% of the 15,051 subjects in Rait et al (2005) had information on the Mini-Mental State Examination (MMSE) missing while in Breeze et al (2004), only 90% of a potential sample of 9,547 individuals was able to provide full information on quality of life. Harris et al (2005) could not provide an answer to the question about whether a nursing-led inpatient unit was more cost effective than usual care because there was inadequate data due to limited follow-up period of the study. In Round et al (2004), there was missing data on outcome data due to patients being excluded, withdrawing or dying. According to Tomassini (2006), refusal rates for older people in general surveys are high but this is not the case in surveys that have been specifically designed for this age group. Up to 17% of the sample in Grundy and Sloggett (2003) had missing information on medicines while Beaumont and Kenealy (2004) had missing data for three out of a selected group of 193 individuals during the first interviews and attrition reduced the numbers to 159 and 143 for the second and third interviews, respectively. Seshamani and Gray (2004a) used only projected hospital expenditures and did not have access to longitudinal cost datasets for nursing home care and general practice.

Beech et al (2004) had non-response rates of up to 40% in their study while Mann et al (2006) had women refusing to take part in their study as it had an opt-in design. Griffiths et al (2000) report that in an earlier pilot study (Griffiths & Wilson-Barnet 1998), there was evidence of selective attrition resulting from patients refusing to participate in the experimental group due to a preference for the control group, a nurse-led inpatient unit. In Griffiths et al (2000), a high refusal rate was recorded with

that in the control group being 23% higher than the one in the treatment group. About 1.4% of the data on patient sex or date of discharge from hospital in Seshamani and Gray (2004b) were missing. In O'Reilly et al (2008), there were missing quality-adjusted life years (QALYs) scores while data on EuroQol EQ-5D (EQ-5D) and Barthel scores were missing for up to 42% of the patients in some cases in Kaambwa et al (2008). Attrition led to bias in the estimation of longitudinal factors such as age, cognitive ability, poor functioning, smoking history, residential status, population mobility and self-perceived health state in Matthews et al (2004). In Walker and Jamrozik (2005), it was not possible to collect all the outcome data. Data on service users' actual involvement and public services policy in the UK were not available in Roberts (2002).

In Downing and Wilson (2005), data on all Accident and Emergency (A&E) attendances were only available for 14 of the 20 acute trust A&E departments in the West Midlands of the UK while some outcomes were not recorded in the trials reported in the systematic review conducted by Parker et al (2000). The sample of disabled people was reduced by 332 because of death in Matthews et al (2005) and information on activities of daily living, income and mean-tested benefits was missing on a number of individuals in the same study. In Fletcher et al (2004), it was not possible to calculate sample sizes for institutional admissions because national data were not available. Clinical outcome data on 24 out of a possible sample of 201 patients were missing in Richardson et al (2001) and Carpenter et al (2003) indicate that data on cognitive impairment were incomplete. In Wilson et al (2003), data on several baseline characteristics were missing.

3.5.2.2 Dealing with missing data

The majority of studies with missing data simply ignored the missing observations (complete case analysis). Individuals with missing values on the MMSE were simply disregarded and excluded from the analysis reported in Rait et al (2005). This is because, the authors submit, there were no clear guiding principles about how to interpret missing values on the MMSE, even though zero was most times assigned to such values. In Grundy and Sloggert (2003), individuals with missing data were excluded from their analysis as was the case in Seshamani and Gray (2004a) who contended that adding costs for nursing home care and general practice would not have negated their result but just modified them. In Matthews et al (2005), individuals with missing data on activities of daily living were also excluded from the analysis and a similar approach was taken in a number of other studies (Beech et al 2004; Mann et al 2006; Round et al 2004; and Seshamani and Gray 2004b).

Other studies did not provide enough information about how they dealt with the missing data. For instance, there was no mention of how the missing data on quality of life was dealt with in Breeze et al (2004).

Other studies acknowledged the problem of incomplete data but do not seem to have done anything to account for this missingness (Beaumont & Kenealy 2004; Walker and Jamrozik 2005; Downing and Wilson 2005; Roberts 2002; Tomassini 2006; Richardson et al. 2001 and Carpenter et al. 2003). Some studies justified not accounting for missing data using results from subsequent analyses that showed that missing data did not lead to bias. A judgement that there was insignificant bias due to missing data on outcomes was made in Round et al (2004) because of similar patterns being identified in the two cohorts compared in the study. In Beech et al (2004),

subsequent workshop discussions about the data collected led the authors to conclude that their findings were not affected by any systematic bias in the data collection while randomisation in Griffiths et al (2000) showed that no significant bias was introduced due to missing data.

Some ‘ad hoc’ approaches were employed in other studies. An additional category in the analysis was created for those with missing data on income and means-tested benefits in Mathews et al (2005). Mean imputation was carried out in O’Reilly et al (2008) while median imputation was conducted in Wilson et al (2003). Imputation refers to an approach where missing values in the variable of interest are replaced by other values obtained through a specified process (Manca & Palmer 2005). Mean imputation involves imputing the mean amount of the variable with missing observations for each missing value. In median imputation, the median is used in place of the mean. Last observation carried forward (LOCF) was also utilised in Wilson et al (2003) where the trial Barthel scores were substituted for the missing Barthel scores during the post-trial period. Only one study, Kaambwa et al (2008), carried out multiple imputation which has been lauded as one of the best methods for dealing with missing data (Schafer 1997; Rubin 1987; Little & Rubin 1987, 2002). In multiple imputation, several different imputed values for each missing observation are created and then using special rules, the results are combined to yield a final result that accounts for the increased uncertainty that arises due to the use of the imputed data (Schafer 1997). This approach is covered in greater detail in section 6.4.2 of chapter six. It is however important to recognise the fact that the criteria used in this literature review has restricted the sample and this may explain why we do not have many studies in our results that have utilised multiple imputation.

What these results seem to suggest is that there is need for quantitative studies that are dealing with populations of older people in the UK to demonstrate the use of multiple imputation when dealing with missing data. The use of multiple imputation is illustrated using the demonstration dataset in chapter seven.

3.5.2.3 Uncertainty of data estimates

This problem was present in 16 studies. In most instances, different methods of estimation of costs and other variables yield different results and it is not always clear which results are better estimates. For costs, this is especially true when there is no universally accepted estimate available and therefore researchers have to estimate costs from resource use data or indeed from budget data. Harris et al (2005) decided to utilise three separate methods for estimating inpatient costs i.e. bottom up costing, bottom-up costing with discounting of the cost of nursing and top-down costing. All the resulting estimates were then used in subsequent analyses. Walsh et al (2005) could not obtain detailed breakdowns of cost for hospital resource units. As a result, cost estimates from different sources were not robust and were also at variance. Specific differences were observed between inpatient and total costs, both calculated from a secondary care perspective. As a result of the uncertainty surrounding medical coverage for, and therefore the cost of, the nursing-led intermediate care unit, Richardson et al (2001) utilised two cost estimates. Similarly in Griffiths et al (2001), three estimates of costs per patient were used (bottom-up costs, discounted costs of nursing and simple cost per bed day). A top-down rather than a bottom-up cost estimate was used in Kaambwa et al (2008) in calculating the cost per intermediate care patient. As a result, there was less between-patient variability introduced in the cost model than would have been the case otherwise. Further, the use of health sector

as opposed to societal perspective costs meant that the contribution from families and the voluntary sector was not captured.

Mann et al (2006) submit that reliance on cross-sectional data may have brought bias into their study and self-report data was also unreliable. O'Shea (2004) found it difficult to estimate the true cost of volunteers' contribution to the senior help line in Ireland. In Grundy and Sloggert (2003), depression could have influenced the score on the self-reported social support measure casting doubt on its accuracy. Fletcher et al (2004) felt that there was a possibility of having underestimated the hospital and institutional admissions in their study while the number of fallers and costs associated could have been under-represented in Newton et al (2006) because of having excluded data from the winter months. The estimates obtained from Cox regression models in Chen et al (2005) were found to be conservative as risk scores used to identify the risk of falls were likely to have been underestimated. Beech et al (2004) based their data on patient case notes which they rightly regarded as having variable quality. The failure to accurately note the population that was eligible to participate in the 'Keep Well AT Home' (KWAH) project in Walker and Jamrozik (2005) led to significant bias. No reliable data was available in Forder and Netten (2000) that could capture information on quality factors like 'atmosphere' and 'staff attitude' among others while Tomassini (2006) underlines the difficulty of comparing in-patient visits to hospital and community care services for the oldest old because of lack of comparable data.

3.5.2.4 Dealing with the problem of uncertainty of data estimates.

To address the problem of uncertainty in estimates, sensitivity analysis was performed in a number of studies (O'Reilly et al. 2008; Harris et al. 2005) where different

assumptions were made. In some studies (Kaambwa et al. 2008), uncertainty in estimates of cost and other variables was simply acknowledged as a limitation with no further action taken to address the problem deemed possible while caution was advised in others (Grundy & Sloggett 2003). In Beech et al (2004), workshop discussions with members of the rapid response intermediate care team as well as health and social care professionals led them to conclude that their results were not affected by systematic bias. In others, strong assumptions were made when discounting uncertainty (O'Shea 2004).

3.5.2.5 Lack of generalisability

Fifteen studies reported this problem. The implication is that results obtained from an ungeneralisable study would not be representative of phenomena in other populations. The results from the cost-effectiveness analysis of a unique early discharge and rehabilitation service (EDRS) in Miller et al (2005) could only be generalised to early discharge services that were similar to the EDRS i.e. not under-resourced, under-skilled or badly managed. In Round et al (2004), non-randomisation of the two comparison groups evaluated was seen as a study limitation as it meant that the two groups were different and the results in one could therefore not be generalisable to the other. The failure to obtain a sufficiently broad range of respondent characteristics despite the use of random sampling in Roberts (2002) meant that no claims to representativeness could be made about the data collected. The results of the systematic review of costs, quality and effectiveness of different locations of care for older people reported in Parker et al (2000) was not generalisable due to the use of a strict inclusion and exclusion criteria.

Forder and Netten (2000) point to the overrepresentation of metropolitan local authorities in their sample as a cause of concern for sample representativeness. Cost data used in Griffiths et al (2005) were specific to local circumstances and could therefore not be generalised very much. Similar observations were made in Fleming et al (2004) and in Oliver et al (2004). In the former, much of the study results were from the Social Services care home rehabilitation service (CHRS) based in Nottingham and the results could not be applicable to CHRSs elsewhere which were staffed and organised differently. In the latter, heterogeneity in settings meant that the risk assessment tools from one setting could not be as effective in another setting or population of patients that was different from the index study. Newton et al (2006) under-estimated the number of falls for individuals as well as the associated costs in a study that examined the costs of falls in Newcastle resulting in ungeneralisable findings. In Mann et al (2006), the study was based on a sample of community-dwelling older women and the results obtained would therefore be difficult to generalise to a larger population that comprised of men too. Further, the former group was significantly different because it comprised a group of women who had all refused to be part of another trial, thereby suggesting potential participant bias. Matthews et al (2005) used a sample from a small town with a predominantly white affluent population and therefore the results from this study could not be generalisable to all of the UK.

3.2.5.6 Dealing with lack of generalisability

One way of avoiding the problem of lack of generalisability is by ensuring that data from a representative sample is used. Tomassini (2006) posits that the nature of a household survey for older people explicitly excludes those living in communal or institutional establishments. To overcome this problem, Tomassini (2006) used data

from the Office for National Statistics (ONS) longitudinal study which includes older people living in communal establishments. Sometimes the problem of lack of generalisability is deemed to be negligible and therefore not requiring remedial action. In Grundy and Sloggert (2003), restricting the age-range in the sample to 65-84 years resulted in a sample that excluded individuals living in institutions. However, the excluded sample reduced the sample by only about 4% which was considered negligible and therefore the problem of generalisability was ignored.

3.5.2.7 Problems associated with distributional characteristics of variables

Another problem, predominantly multivariate in nature, was that linked to the distributional characteristics of the (dependent) variables and was reported in 14 studies (Miller et al. 2005; Kaambwa et al. 2008; Round et al. 2004; Newton et al. 2006; Griffiths et al. 2000; Seshamani & Gray 2004b; Forder & Netten 2000; Griffiths & Wilson-Barnet 2000; Griffiths et al. 2001; Gunnel et al. 2000; Cotton et al. 2000; Shaw et al. 2000; Young et al. 2007 and Green et al. 2005). In Forder and Netten (2000), the Breusch-Pagan chi-square test confirmed the presence of heteroscedasticity² in the error terms of the ordinary least squares (OLS) regressions that were run while multicollinearity was found in the Heckman sample selection model reported in Seshamani and Gray (2004b). Heteroscedasticity leads to inefficient estimators and biased standard errors while multicollinearity leads to imprecise estimates (Gujarati, 1995). Most cost data used in analysis were skewed as shown in Miller et al (2005), Newton et al (2006), Kaambwa et al (2008) and Seshamani and Gray (2004b). Skewness or non-normality is asymmetry in the distribution of the sample data values (Altman 1991). Positive (negative) skewness is when a greater proportion of the observations are less than (greater than) the mean implying that the

² Heteroscedasticity occurs when the variance of the OLS error terms is not constant (Gujarati 1995).

mean is larger (smaller) than the median (Altman 1991). Means and standard deviations from skewed data will not be reliable. Most cost distributions are positively skewed because of the presence of a lot of zero cost observations (Seshamani & Gray 2004b). Other variables were also shown to suffer from non-normality as was the case in Round et al (2004) where the EuroQol weighted health index scores were skewed and in Griffiths et al (2001) where the Barthel index was not normally distributed. Other studies also had skewed variables (Gunnel et al. 2000; Cotton et al. 2000; Shaw et al. 2000 and Green et al. 2005).

3.5.2.8 Dealing with problems associated with distributional characteristics of variables

Several approaches were used to address the problem of skewness. Using recommendations of others like Barber and Thompson's (2000), bootstrapping was used in comparisons of mean costs in the ensuing cost-effectiveness analysis in Miller et al (2005) and in Griffiths et al (2001). Bootstrapping enables the estimation of population measures without having to make assumptions about the distribution (Briggs & Gray 1999). Seshamani and Gray (2004b) and Kaambwa et al (2008) employed a generalised linear model (GLM) framework that used a log link as advocated for by Manning and Mullahy (2001). In Round et al (2004) and Griffiths et al (2001), non-parametric tests were used to compare EuroQol health index and Barthel scores, respectively, while in Griffiths and Wilson-Barnett (2000), the length of stay variable was subjected to natural log transformation so as to comply with underlying regression assumptions. Whilst acknowledging the problem of skewness, other studies such as Newton et al (2006) did not do anything about it submitting that the results were not biased by this skewness.

Robust t-ratios were used to overcome the problem of heteroscedasticity in Forder and Netten (2000) while the use of a two-part regression model and additional regressors was used to address the problem of multicollinearity in Seshamani and Gray (2004b).

What the studies reviewed in this section seem to suggest is that the challenge of distributional characteristics of a (dependent) variable is one that has been overcome in many studies dealing with service evaluations of older people. There are therefore a number of examples of good practice that one can draw upon when dealing with a similar population and comparable statistical issues.

3.5.2.9 Problems of sample size and lack of power

Ten studies were shown to have this problem and the result would be that unreliable estimates would be obtained (Altman 1991). In Tomassini (2006), as in many other surveys, the sample size of old people was small while Walsh et al (2005) posit that using bottom-up costing method would have resulted in a small sample size because of restrictions that arose due to ruling out unreasonable resource input cost estimates. In Breeze et al (2003), explanatory models investigating determinants of quality of life in old age in Britain were not run because the cells were too small. As a result of the limited number of respondents in Matthews & Brayne (2005), assessment of incidence of dementia was not possible while the low numbers in Beaumont and Kenealy (2004) also made the evaluation of the impact of different strategies difficult. The number of trials in Parker et al (2000) was small for some intermediate care services such as nurse-led beds and hip care units and the authors point out that this meant that no firm messages about these services could be obtained from their study.

3.5.2.10 Dealing with problems of sample size and lack of power

The obvious solution in cases where there is a problem of small sample sizes/lack of power is to increase the sample size. To reach an acceptable sample size in Tomassini (2006), different series of cross-sectional data were combined while top-down costs were used so as to arrive at a reasonable sample size in Walsh et al (2005). Where this was not possible, caution was advised when interpreting the results obtained (Richardson et al. 2001).

3.5.2.11 The need for predicted measures of outcome.

Another problem identified, and related to that of missing or incomplete data, is the need to predict one outcome measure from another if the former had not been collected. This was seen in 6 studies. In particular, there was need to predict a utility based measure from a non-utility measure where the former was missing. This issue was reported by five studies (Rait et al. 2005; Matthews & Brayne 2005; Mathew et al. 2004; Harris et al. 2005 and Tomassini 2006). When patients self-report data, it is possible that the data may be inaccurate because of reporting bias or because of the inability of patients to provide information about themselves (Mann et al. 2006). The latter is common for older people who may not be mentally or physically able to complete questionnaires if they have conditions such as dementia (Matthews & Brayne 2005), cognitive impairment (Rait et al. 2005) or depression (Grundy and Sloggett, 2003). In the Harris et al (2005) study, the average age of patients was 78 years. These patients were not capable of filling out the Barthel Index scale and as such, proxies, who in this case were researchers and nurses involved in the direct care of patients, completed the scale. Interviews that could not be completed by patients in another study (Matthews et al. 2004) were also completed by a proxy. The interviews involved the completion of outcome measures such as the Mini- Mental Examination,

ADL and the Townsend Activities of daily living score and the Geriatric Mental State (GMS) (Matthews et al. 2004). The use of proxies can lead to interviewer bias which according to Matthews & Brayne (2005) has the potential of influencing the results especially if it is done consistently. Matthews & Brayne (2005, p.760) controlled for this by conducting “...systematic training, regular quality control and highly structured interviews.”

3.5.2.12 Dealing with the need for predicted measures of outcome.

Tomassini (2006) shied away from using a self-rated general health status question for older people opting for one that asked about the presence of a long-standing illness to avoid having to use proxy interviewees. Rait et al (2005) also proposed the use of other instruments other than the self-reporting MMSE that was used in their study. In Forder and Netten (2000), dummy variables were used to account for the lack of appropriate proxy variables but these were later dropped on statistical, theoretical and parsimony grounds. Diagnostic tests showed that this omission of the dummy variables did not lead to misspecification.

What are lacking from these studies are examples of analyses that have predicted one outcome measure from another when the former is missing. When data on one outcome measure for individuals are missing but information has been collected on another measure of outcome for the same individuals, it may be possible to map the latter onto the former. This has been done in studies elsewhere. For instance, in a study that examined a sample of 598 stroke patients in the Netherlands (van Exel et al. 2004), individuals' missing values on the EQ-5D were predicted from Barthel Index scores using a regression model framework. The adoption of this method of dealing with missing measures of outcome is the thrust of Chapters eight and nine. The results

will be helpful in informing researchers on the options available when important measures of outcome needed for other analyses such as economic evaluation are missing.

3.5.2.13 Other statistical issues identified

Other problems recognised included lack of causality (found in four studies) and numerous other forms of bias. Lack of causality refers to situation where one can not attribute the occurrence of one variable of a certain class on the occurrence of another variable of another class as was the case in Breeze et al (2004) or in Grundy and Sloggert (2003). A common solution to this would be to re-specify the models displaying problems of causality so as to incorporate variables with a cause-effect relationship; distinguish between causality and other relationships such as correlation (Gujarati, 1995). The other forms of bias reported were diverse and included, among others, participant and response bias (Mann et al. 2006), selection bias (Grundy and Sloggert, 2003), heterogeneity in data collection periods (Newton et al. 2006) and censoring (Seshamani and Gray, 2004b).

3.5.3 Links between the statistical problems identified in studies

Some of the statistical problems discussed in section 3.5.2 are not mutually exclusive. It is possible for one problem to arise because of another and vice-versa. Further, it is possible for more than one problem to be present in a single dataset.

Problems with the distributional characteristics of the dependent variables may be as a result of the type of sample chosen. If the sample is atypical, for instance, then there may be problems with skewness e.g. having a higher proportion of older people in a sample will inevitably lead to positively skewed health costs as the majority in this

kind of sample will have low health costs as was the case in Newton et al (2006). Heteroscedasticity can be caused by measurement error (Gujarati 1995) which also leads to uncertainty in data estimates. Heteroscedasticity may also arise due to outliers (values which are not representative of the rest of the observations) especially if the sample is small (Greene 1997). The presence of multicollinearity (where two or more independent variables in a regression model are linearly related to each other (Gujarati 1995) will lead to coefficients that cannot be estimated with precision or accuracy. Another problem may be autocorrelation (“...correlation between members of a series of observations ordered in time or space” (Kendall & Buckland, 1971, p.8)). It can also be the result of ‘exclusion bias’ which may result from missing data (Gujarati 1995).

The need to predict missing measures of outcomes may arise because of missing data on the preferred outcome measures (van Exel et al. 2004) and may also be because of inaccuracies in the way the measure was calculated thereby leading to uncertainties concerning their accuracy (Mann et al. 2006).

Small sample sizes may arise because data is missing and therefore individuals have to conduct complete case analysis (Grundy & Sloggett 2003). Another cause may be uncertainty in the estimates obtained for some variables and therefore ignoring such variables may seem a logical step which action ultimately leads to a small sample.

Lack of generalisability will most often be linked to the sample used. The problem may be that the sample used is atypical and therefore the findings would not hold for a sample with different demographics (Mann et al. 2000; Fleming et al. 2004). It may also be because the sample chosen was just small and therefore did not have as much

variation in its composition compared to the total population. In addition, this problem may arise because there is a lot of missing data which results in a small sample or in a sample that has only certain types of data missing thereby resulting in an unrepresentative dataset. If uncertain or inaccurate estimates were used in an analysis, then it is also possible that the results may not be generalisable

Uncertainty in estimates may come about because some information is missing and therefore the calculation of certain measures is inaccurate. It may also be due to unique distributional characteristics of some variables which make certain estimates incorrect. For instance, the use of the mean rather than the median for summarising skewed data will lead to inaccurate inferences being drawn (Altman 1991). Another source of inaccuracy is the use of wrong measures of an outcome when the correct one is missing. For example, relying on a measure of functional independence such as the Barthel index (Mahoney & Barthel 1965) to gauge health status when health status measures are missing would be inaccurate. This problem can be compounded further by the unavailability of other measures of outcome that can be used as proxy measures or to predict the missing measure. The method used to predict the missing outcome measure from another, where this is possible, may also lead to uncertainties in the estimates ultimately obtained. This latter issue is examined in more detail in chapters eight and nine.

3.5.4 Key messages for health services researchers from this chapter

Researchers dealing with quantitative datasets obtained from evaluations of services for older people need to be aware of a number of potential statistical problems that can be present in such datasets. Table 3.3 is useful for identifying the common statistical issues one has to be on the look out for when dealing with such datasets.

These problems, in order of prevalence in published evaluations in the UK, included missing data, lack of generalisability, problems associated with the distributional characteristics of variables, sample size and lack of power problems as well as the need for predicted outcome variables. Other statistical problems are lack of causality, different types of biases (participant, response and selection biases), problems due to heterogeneity and censoring. When these problems are present in a dataset, it may not always be possible to carry out ‘straight-forward’ analyses. Various methods for dealing with these statistical problems, each with different assumptions and/or underlying theories have been presented in this chapter and it will therefore be important for researchers to be aware of these. There is however scope for other statistical methods used in other populations to be used or adapted to datasets drawn from populations of older people. Further, where more than one statistical problem exists, it will be important for researchers to determine if there is any link of causality between the problems as such links may have a bearing on the most appropriate method for tackling the problems.

3.6 Conclusion

The preceding review was carried out in a bid to identify the common statistical challenges that can be found in quantitative studies evaluating services for older people in the UK. A summary of these challenges is provided in Table 3.3. The resultant objective of this exercise was establishing what methods have been used to overcome these problems and whether the methods could be applied to datasets with similar characteristics. In particular, methods relevant to the statistical problems also found in the demonstration dataset were sought. A number of statistical problems were identified but it is important to point out that there are more problems that may not have been revealed during this exercise. Amongst the most common problems that

researchers in this or other related populations need to be aware of were missing data, lack of generalisability, problems associated with the distributional characteristics of variables, inaccuracies or unreliability of variables, small sample sizes/lack of power and the need for predicted measures of outcomes.

A number of solutions to tackle statistical problems in datasets obtained from populations of older people, which can be used by researchers dealing with such datasets, were highlighted in the studies reviewed.

In dealing with missing data, simple methods such as ignoring missingness or mean imputation were used in the studies reported. Only one study used multiple imputation. Therefore, there is scope to demonstrate how this problem can be overcome using principled or much more robust methods. In chapter seven, multiple imputation methods are further compared to complete case analysis and the Heckman sample selection method in a regression framework.

Sensitivity analysis has been shown to be a principled way of dealing with the problem of uncertainty in estimates. Other ways of ensuring that estimates used in an analysis do not lead to bias have also been shown.

Using data from representative samples is the best way of addressing the problem of lack of generalisability as was shown in the studies reported. Where this is not possible, care should be taken not to generalise the results of an analysis where this problem is cause for concern.

A number of studies had used principled methods to deal with problems associated with distributional characteristics of the variable. Remedies used included modelling (e.g. the GLM for non-normality, use of robust t-ratios for heteroscedasticity, transformation of variables or the use of a two-part regression model for multicollinearity). These results therefore provide scope for the adaptation of some of these methods in the analysis of the demonstration dataset. In chapter five, some problems that are linked to the distributional characteristics of variables are addressed using a GLM framework.

The need for predicted measures of outcomes is the third common problem considered. Some studies have used proxy respondents as opposed to proxy measures of outcomes while some have suggested that the solution lies in changing the design of studies e.g. removing questions where respondents self-report their status. These results therefore present another opportunity to demonstrate how this problem can be dealt with by predicting one outcome measure from another. This is done in Chapter nine.

The next chapter describes the national evaluation of costs and outcomes of intermediate care services for older people in the UK. This national evaluation provided the demonstration dataset that has been used in the empirical analyses reported in chapters five, seven and nine. The chapter also discusses the statistical problems that were identified in the demonstration dataset, which problems were the subject of the empirical analyses.

Figure 3.1: Flow Chart of Literature Review

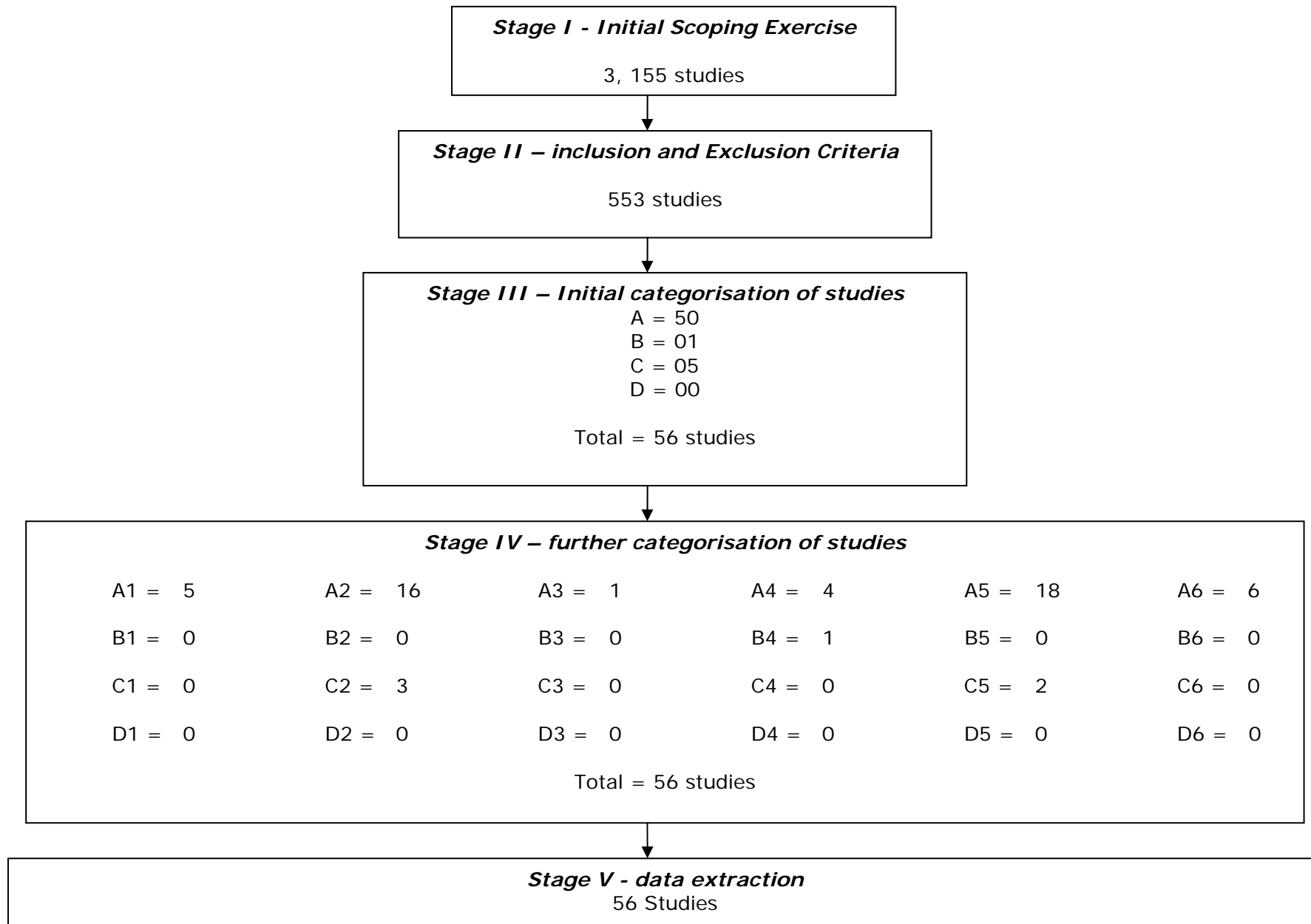


Table 3.1: Studies excluded at Stage IV

Study	Final Class	Primary Focus	Reason for exclusion
Bago-d'Uva (2005)	E	Modelling access to and utilisation of primary care using data from the British Household Panel Survey 1991-2001	Sample not exclusively for older people
Bajekal et al (2004)	D	Investigating influence of ethnic differences on quality of life at older ages	Sample is of respondents aged 45-74 years
Beale et al (2005)	E	Comparisons the cost of care in general practice – patients compared by council tax valuation band of their home address	Sample not exclusively made up of older people
Blackburn et al (2005)	E	Cross-sectional survey of adult carers n their use of the internet	Focuses solely on carers
Britton et al (2004)	E	Determining whether access to cardiac procedures and drugs contributes to social and ethnic differences in coronary heart disease in a population setting	Sample aged 35-55 years
Burholt (2004)	D	An assessment of the settlement patterns of and residential histories of older Gujaratis, Punjabs and Sylhetis in Birmingham, England	Older people aged 55 years or older considered
Casado-Diaz et al (2004)	D	Evaluation of characteristics, motivations and adjustment of northern European retired residents	Not UK based sample
Department of Health (2005)	E	Examination of investment levels in adult mental health services in England for 2004 to 2005	Not specific to older people
Drever et al (2004)	E	Examining the relationship between class, gender and self-rated health in adults in Great Britain	Participants aged 25-64 years
Ebrahim et al (2004)	D	Analysis of social inequalities and disabilities in older men	Men aged 52-73 years considered
Ellis et al (2006)	D	Cost-effectiveness analysis of a joint NHS/Social Services short-term residential rehabilitation unit compared with 'usual' community services at home for older people at discharge from hospital	Patients aged 55 years or older considered
Evandrou (2000)	D	Investigating inequalities in later life for ethnic minority groups in Britain	People aged 60 years and over assessed
Evans et al (2003)	E	Survey of psychiatric morbidity of older people	Sample includes ages 16 – 74 years
Forster et al (2008)	D	Medical day hospital care for the elderly versus alternative forms of care	Sample includes patients aged 60 years
Gilbertson et al (2000)	D	Comparison of domiciliary occupational therapy to usual care for stroke patients discharged from hospital	Sample included patients as young as 28 years
Godfrey & Townsend (2008)	D	Examination of the experience of illness and process of recovery for intermediate care patients	Samples includes patients aged 64 years
Green et al (2005)	E	Assessment of the clinical and cost-effectiveness of drotrecogin alfa (activated_) for treatment of adults with severe sepsis in a UK context	Focus on patients under 18 years of age
Griffiths et al (2005)	D	Post-acute intermediate care in nursing-led units: a systematic review of effectiveness	Patients as young as 18 years old were considered.
Griffiths et al (2007)	D	Assessing the effectiveness of intermediate care in nursing-led in-patient units (NLIUs) compared to usual care	Patients from 18 years old were considered.
Hardy & Kuh (2005)	C	Assessment of whether socio-economic status, environmental stress and hardship throughout the life course are associated with age at menopause	Sample not old enough i.e. middle aged women (47 – 53 years) was considered

Table 3.1: Studies excluded at Stage IV

Study	Final Class	Primary Focus	Reason for exclusion
Healey et al (2004)	D	Testing the efficacy of a targeted risk factor reduction core care plan in reducing risk of falling while in hospital	Older people aged 63 years and over in the sample
Hyde et al (2000)	D	Systematic review to investigate the effects of supported discharge after acute admission in order people with undifferentiated clinical problems	Includes non-UK studies
Mold et al (2003)	E	A review of and commentary of social factors which influence stroke care	Qualitative study
Norman et al (2005)	E	A longitudinal study analysis of selective migration, health and deprivation in England and Wales	Sample not specific to older people
Parker et al (2002)	D	A systematic review of discharge arrangements for older people	Included non-UK studies as well
Perry & McLaren (2004)	D	Exploration of nutrition and eating disabilities in relation to quality of life at 6 months post-stroke	Sample not entirely made up of older people
Power et al (2005)	D	Investigation of the contribution of childhood and adult socio-economic position to adult obesity and smoking behaviour	Participants were aged 30-50 years and survey also included Denmark, Finland. The Netherlands, Sweden and the USA
Roderick et al (2001)	D	Comparison of the effectiveness and costs of a new domiciliary rehabilitation service for elderly stroke patients with geriatric day-hospital care	Stroke patients aged 55 years or over included
Roe et al (2003)	D	Evaluation of client information across Intermediate Care schemes	Older people aged between 50 and 101 years have been included in sample
Scuffham et al (2003)	D	Estimation of costs of different fall types and by age group	Sample is made up of people aged 60 years or over
Shepperd and Illife (2005)	D	Comparing the effects of hospital at home (HAH) with those of in-patient hospital care	Included people aged as young as 18 years as well as non-UK studies
Smith et al (2004)	D	Evaluation of factors that determine quality of life among older people in deprived neighbourhoods	People aged 60 and over included in sample
Spiers et al (2005)	D	An investigation of whether standard of living predicts health decline in older people as well as a comparison of standard of living with other indicators of socioeconomic status	Age range of older people is between 55-69 years
Steiner et al (2001)	D	Comparison of post acute intermediate care in an inpatient nurse-led unit with conventional post-acute care on general medical wards of an acute hospital	Sample includes patients at least 16 years old
Thomas et al (2005)	E	Examination whether patients with persistent non-specific low back pain gained more long-term relief from pain when offered access to traditional acupuncture care alongside conventional primary care than those offered conventional care only	Age group considered was 18-65 years
Young et al (2005a)	D	Evaluation of clinical outcomes of a new city-wide Intermediate Care service and its effects on hospital and institutional use compared to usual care	Frail older people aged 63 years or older considered
Young et al (2005b)	D	Study of baseline characteristics of patients presenting to two elderly care departments as emergencies with the clinical syndromes of falls, incontinence, confusion or poor mobility	Frail older people aged 63 years or older considered

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Harris et al (2005)	A1	Evaluate costs from the UK National Health Service perspective, of transfer to a nursing-led inpatient unit for intermediate care	Patients with mean age 78 years in NLIU arm and 79 years in control arm	175	Randomised Clinical Trial (Cost-effectiveness Analysis)	UK nursing-led Inpatient Unit	Barthel Index, Cost per Day, length of stay	Nursing-led inpatient units have higher costs of hospital stay (despite having a lower cost/day) than acute hospital wards because of longer length of stay in the former.
Miller et al (2005)	A1	To calculate the cost-effectiveness of an early discharge and rehabilitation service (EDRS) compared to usual care.	Patients aged 65 or over	370	Randomised Clinical Trial (Cost effectiveness analysis)	Acute and rehabilitation wards in NHS hospitals in Nottingham	Costs, EuroQol EQ-5D, Quality adjusted life years (QALYs)	The Nottingham EDRS was more cost effective than usual care.
O'Reilly et al (2006)	A1	Cost effectiveness study of post-acute care for older people in a community hospital compared with care in a district general hospital department for older people with acute medical conditions	Older people aged at least 76 years	220	Randomised Clinical Trial (Cost-effectiveness Analysis)	Community hospital and district general hospital in Yorkshire, UK	Costs, EuroQol EQ-5D, Quality adjusted life years (QALYs)	There is no difference in cost-effectiveness between the two groups.
O'Reilly et al (2008)	A1	Comparison of the cost-effectiveness of post-acute care for older people in community hospitals and general hospital care	Older people aged 65 and over	490	Randomised Controlled Trial (cost-effectiveness study)	Seven community hospitals and five general hospitals at five centres in the midlands and north of England.	EuroQol EQ-5D and health and social service costs	The cost-effectiveness of post-acute provided in community services was similar to that provided in general hospital care.
Walsh et al (2005)	A1	Economic evaluation of nurse-led intermediate care compared with standard hospital care for post-acute medical patients	Older patients – age range not reported	238	Pragmatic Randomised Clinical Trial (Cost-minimisation analysis from an NHS perspective)	Nurse-led unit and acute general medical wards in a large urban UK teaching hospital	Costs	Nurse-led intermediate care more expensive than standard hospital care.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Boston et al (2001)	A2	Comparison of elderly patients admitted to an inner city general practitioner (GP) unit with comparable patients in conventional care	Older people aged 65 and over	228	Prospective non-randomised comparative study	An Inner London Health Authority	Demographic and medical data; cognitive function (Abbreviated Mental Test); mental state (Philadelphia morale scale); SF-12; quality and quantity of support	Physical outcomes in GP units and conventional settings were similar. GP units were associated with short-term improvement in mental functioning and better quality of care.
Breeze et al (2004)	A2	Identification of socioeconomic differentials in quality of life and their explanatory variables among older people	People aged 75 and over	9, 547	Cluster Randomised Controlled Trial (Poisson Regression without a time element)	23 General practices in Britain (Medical Research Council – MRC Trial of Assessment and Management of Older People in the Community)	Sickness impact profile (SIP), and the Philadelphia geriatric morale scale	Poor quality of life was associated with: renting rather than owning a home; self-reported health problems; smoking; alcohol consumption; low socioeconomic position; living arrangements; poor home management and poor morale.
Cotton et al (2000)	A2	Comparison of early discharge with home treatment of patients supported by respiratory nurses with usual hospital care	Older people aged at least 65 years	81	Randomised controlled trials	Glasgow Royal Infirmary, UK	Readmission; additional hospital days, deaths within 60 days of initial admission	Early discharge policy was associated with reduced inpatient stay. There was no difference in the subsequent need for readmission between the two groups.
Davies et al (2000)	A2	Comparison of 'hospital at home' and hospital care as an inpatient for patients with an exacerbation of chronic obstructive pulmonary disease	Older people with mean age of 70 years	150	Prospective randomized controlled trial	University teaching hospital, Liverpool, UK	Readmission rates; changes in forced expiratory volume in one second (FEV ₁); Mortality	Hospital at home was found to be a practical alternative to usual care for certain patients.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Fleming et al (2004)	A2	Assessment of the effect that a care home rehabilitation service has on institutionalisation, health outcome and service use	Elderly and disabled people aged over 65 years	165	Pragmatic Randomised controlled trial (Contingency table analysis, multiple linear regression, non-parametric tests)	Patients living in the Social Services districts served by the scheme who also wished to go home but were at high risk of being institutionalised.	Institutionalisation rates; Barthel index; Nottingham Extended ADL score; General Health Questionnaire; Health and Social Service resource use	There was no reduction on institutionalisation due to the service compared to usual care.
Fletcher et al (2004)	A2	Measurement of the effects of different approaches to assessment and management of older people	Patients 75 years and over	43, 219	Cluster-Randomised factorial Trial (Cox's proportional hazards, Poisson & logistic regression analyses)	106 UK practices participating in the MRC trial	Mortality, hospital and institutional admissions; quality of life (sickness impact profile & the Philadelphia geriatric morale scale)	There was no difference between groups in terms of institutional admissions, mortality and hospital. Positive association between quality of life & homecare (universal vs. targeted assessment) and mobility, social interaction and morale (geriatric vs. primary-care team).
Green et al (2005)	A2	Determination of the effect that a locality based community hospital has on independence in older people needing rehabilitation compared to usual care in a hospital ward	Older people with mean age of at least 86 years	220	Randomised controlled trial	Community hospital and district general hospital in Bradford, UK	Nottingham extended activities of daily living scale; general health questionnaire 28; Barthel Index; Nottingham health profile.	There greater association between independence and care in a locality community hospital compared to care in a district general hospital.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Griffiths et al (2000)	A2	Assessment of the potential for a nursing-led inpatient unit (NLIU) – comparison between plan to transfer to NLIU and plan to remain under usual care	Older people with mean age of 77 years	177	Randomized controlled trials (Contingency table analysis, non-parametric tests and analysis of covariance)	An NLIU in an Inner London Hospital Trust	Barthel Index; Length of stay; discharge destinations; mortality; nurse-patient ratios	No significant difference in Barthel index between NLIU and usual care. NLIU associated with longer length of stay and lower nurse-patient ratios.
Griffiths et al (2001)	A2	Determination of the outcome and cost of transferring patients to a nursing-led inpatient unit for intermediate care compared to usual hospital care.	Older people with a mean age of 78.3 years	175	Randomised controlled trial (Contingency tables analysis, Analysis of covariance)	Hospital wards in an acute inner London National Health Service Trust	Length of stay; costs; discharge destination; Barthel Index	NLIU associated with longer length of stay, lower daily cost and higher mean costs. NLIU had no significant effect on discharge destination.
Gunnel et al (2000)	A2	Assessing the impact of early discharge hospital-at-home scheme (compared to hospital) on carer's strain and quality of life	Older people aged at least 71 years	80	Randomised controlled trial	Bristol, UK	Modified 12-item Carer Strain Index; COOP-WONCA charts; EuroQol EQ5D	There was no evidence of an increase in self-reported strain on carers of patients discharged from hospital.
Harris et al (2007)	A2	Determination of whether transfer to a nursing-led inpatient unit (NLIU) prior to discharge from hospital can improve clinical outcome and reduce length of stay and readmission rate for medically stable post-acute patients assessed as requiring inpatient care.	Older people	471	Pragmatic randomized controlled trials (Retrospective secondary data analysis)	Three purposefully replicated, pragmatic randomized controlled trials	Barthel Index; length of stay; discharge destination; mortality; General Health Questionnaire; Nottingham Health Profile Distress Index (NHPDI) ; incidence of complications and readmission	Post-acute patients with complex health and social needs transferred to a NLIU can have better outcomes of care (Barthel; General health questionnaire, NHPDI; pressure ulcers) but longer length of stay.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Richardson et al (2001)	A2	Assessment of the cost and impact on outcomes of introducing a nursing-led ward program	Older people with mean age of 77 years	177	Randomised controlled trial (Effectiveness analysis)	Nursing-led intermediate care unit (NLIU) in an inner London teaching hospital	Length of stay; Costs; Barthel Index; Mortality	No significant difference in outcomes of NLIU and standard care. Costs for NLIUs are higher because of longer length of stay.
Shaw et al (2000)	A2	To consider the effectiveness of multifactorial intervention after a fall in order patients with cognitive impairment and dementia attending the accident and emergency (A&E) department	Older people aged at least 65 years	274	Randomised controlled trial	Two A&E departments, Newcastle upon Tyne, UK	Number of fallers; number of falls; time to first fall; injury rates; Mortality; ; fall related hospital admissions fall related attendances at A&E department	Multifactorial intervention was not effective in preventing falls in older people.
Wade et al (2003)	A2	Assessing whether a programme of multidisciplinary rehabilitation and group support is associated with sustained benefit for people with Parkinson's disease or their carers	Older people with mean age 71.3 years	144	Randomised controlled (crossover) trial	Oxford, UK	EuroQol EQ-5D; SF-36; Parkinson's disease disability questionnaire; Parkinson's disease questionnaire ; hospital anxiety and depression scale; limited stand-walk-sit test; carer strain index	A short spell in multidisciplinary rehabilitation may improve mobility (stand-walk-sit test) but this service had worse general and mental health (SF-36).
Young et al (2007)	A2	Investigation of the effect of delayed community hospital transfer on outcome	Older people with mean age of at least 86 years	220	Randomised controlled trial	Community hospital and district general hospital in Bradford, UK	Time to transfer; Nottingham extended activities of daily living scale	Shorter time to community hospital transfer was associated with improved independence.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Newton et al (2006)	A3	Quantifying immediate cost of attending to fallers	Older people aged 65 and over	1,504	Non-experimental analytic study (Prospective analysis of Service data /Routinely collected data - Cost analysis)	Data from Newcastle collated by the North East Ambulance Service (NEAS)	Number of falls and costs	It costs £146 per fall and ambulance crews spend 2.25 days/month of their time on falls.
Forder and Netten (2000)	A4	Assessment of the impact of (i) production costs (ii) competition and (iii) contract choices on the price of placements of elderly people in residential and nursing care homes	Older people Age-range not given	1780 publicly funded placements	Non-experimental analytic study (Regression analysis of routinely collected data)	Admissions to residential and nursing home care from 18 Local authorities in the UK	Costs, market competitiveness, contract choices	There was a significant association between contract payment arrangements and placement prices.
Seshamani and Gray (2004a)	A4	Construction of a projection model that considers the impact of demographic change on health care costs towards the end of life	Patients aged 65 and older	90, 929	Non-experimental analytic study (Two step Probit Regression analysis of Longitudinal dataset)	Oxford Record Linkage Study (ORLS) and cost data from the UK department of Health	Costs, Age-specific mortality rates	Decline in age-specific mortality rates over time pushes back death-related costs. Accounting for this reduced predicted annual growth rate expenditure by half to 0.40%.
Seshamani & Gray (2004b)	A4	Replication of the Zweifel et al (1999) model showing the impact that age and proximity to death have on health care costs	Older people aged 65 years or over	95, 900	Non-experimental analytic study (Heckman sample selection model; Two-part econometric analysis – Cohort study)	Oxford Record Linkage Study (ORLS), Oxfordshire, England	Cost predictions; probability in hospital and proximity to death	Correct model selection is important for accurate analysis of determinants of health-care expenditure. Two-part model better than the two-step Heckman model.
Ward et al (2005)	A4	Evaluating the effectiveness of a Respiratory Intermediate Care Team (RICT)	Older people	88	Non-experimental analytic study (Audit & patient satisfaction postal questionnaire)	A respiratory intermediate care team (RICT) in Oxford City Primary Care Trust	Costs; Reason for referrals, source of referral; patient satisfaction views	The RICT is effective in saving hospital bed days through prevention of admission and early discharge of chronic obstructive pulmonary disease (COPD) patients. Hospital at home was more favourable. The RICT led to cost savings on hospital admissions.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Beaumont and Kenealy (2004)	A5	A study of the influence that the type of residence and social comparison strategies of healthy older people's have on their quality of life perceptions	Older people aged 65 or over	190	Non-experimental analytic study (Simple statistical analysis of Assessment Interviews)	London Borough of Wandsworth	Schedule for Evaluation of Individual Quality of life: Direct weighting and Social Comparison strategies (Upward identification, Upward contrast, Downward identification & Downward contrast)	There is a significant relationship between quality of life and the most dominant of the social comparison strategies: downward contrast.
Beech et al (2004)	A5	Evaluation of a multidisciplinary Rapid Response Team (RRT)	Patients with mean age 75.9 years	231	Descriptive study (Simple statistical analysis)	Patient care notes & staff, local health and social care providers in Herefordshire Primary Care Trust	Views on the new service, number of service users, medical and social needs	On the whole, patients and carers were positive about RRT but there also disquiet about the lack of influence over choice of care and the quality of information about care.
Brooks et al (2003)	A5	Explore if Intermediate Care could reduce number of unnecessary emergency hospital admissions	Older people aged 65 years or over	57	Non-experimental analytic study (Simple statistical analysis)	Patient discharge summaries (6 July – 30 November 2001) from the Rapid Assessment Support Service (RASS)	Discharge destinations; readmission to hospital; patient needs	The RASS appeared to have prevented unnecessary admissions to hospital for older people. Only 5% were readmitted to hospital.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Carpenter et al (2003)	A5	An evaluation of intermediate care services for older people by examining the relationship between different intermediate care services (ICSSs) and the use of hospital beds and intermediate care resources	Older people aged 65 years and over	1,167	Non-experimental analytic Study (Computer simulation model)	interRA/Minimum Data set for Home Care Assessment (MDS-HC) for Shepway District of Kent, UK	Staff and bed numbers; admission criteria; resource use; Patients' medical, physical, psychological and social characteristics	Patients admitted to ICSSs had different characteristics and some of these did not seem to meet the criteria. Rehabilitation wards were under pressure from high patient demand while the cognitively-impaired were denied access to ICSSs. There was need to increase the capacity in the community-based ICSSs as well as a well-defined admission criteria for ICSSs.
Downing and Wilson (2005)	A5	Description of A & E attendance patterns of older people	Older people aged 65 and over	514, 420	Descriptive study (Simple statistical analysis)	The A & E Commissioning Data Set collected for 14 Acute Trusts in the West Midlands (1 April 1999 to 31 March 2002)	Attendance rates	Attendance rates were highest in those aged over 80 years and in winter. Older patients are more likely to attend in the morning and early afternoon, in winter months, arrive by ambulance and require admission to hospital.
Griffiths and Wilson-Barnet (2000)	A6	Assessment of what influences length of stay in nursing-led inpatient units (NLIUs)	Older people with mean age above 76 years	296	Non-experimental analytic study - (Analysis of covariance analysis)	NLIU in an acute hospital trust that spread over two sites	Length of stay;	Length of stay appears to be associated with the NLIU's location, staffing levels and patient population.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Grundy and Sloggett (2003)	A5	Analysis of variations in the health of older people – Assessment of association between explanatory domains and outcome indicators and what effect explanatory variables have on health status indicators	Older people aged 65-84 years	9, 129	Non-experimental analytic study (Logistic regression analysis of routinely collected data)	1993-95 Health Survey for England	Presence of a long-standing illness, number of specific long-standing conditions, self-reported general health, General Health Questionnaire, Height, education, marital status, age, smoking, sex, social support measure, tenure & income support	Social resources (marital status and social support) were related to psychological health (CHQ), and self-rated health. Smoking more related to these indicators than to self-rated health.
Institute of Health Sciences and Public Health Research (2005)	A5	Examining the structure, process, outcomes and cost effectiveness of intermediate care for older people.	Older people aged 65 and over	7,452	Non-experimental analytic study (Comparative case study)	Primary Health Care trusts from North England	Costs; performance indicators; Service delivery, culture and behaviour data; User experiences and outcomes	Intermediate care is conceived and implemented in diverse ways. High level intermediate care performance indicators did not offer either positive or negative evidence. Intermediate care has however been associated with substantive changes in the structure of service delivery and in the culture and behaviour of commissioners and providers.
Jenkinson & Ford (2006)	A5	Comparison of changes in stroke services in sites that had followed the National Clinical Governance Support Team's (CGST) programme for stroke and in those that did not.	Not reported	200 & 240	Non-experimental analytic study (The Review, Agree, Implement, Demonstrate (RAID) model)	Delegates from 42 Trusts who completed four waves of the CGST Stroke Programme and lead clinicians for stroke in 240 sites that participated in the National Sentinel Audit of Stroke 2001/02	Patient & carer participation in treatment programmes, staffing levels, new funding ,user involvement	There is a significantly greater change for sites that participated in the programme in terms of stroke unit provision, staffing levels and new funding.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Kaambwa et al (2008)	A5	Evaluating costs and outcome of intermediate care patents	Older people aged 65 or over	2,253	Non-experimental analytic studies (Generalised Linear Model Regression methods)	Five anonymous UK case studies of 'whole systems'	Costs; Barthel Index; EuroQol EQ-5D	Almost 50% of patients were inappropriately admitted to Intermediate care. In comparison to supported discharge, admission avoidance services were associated with both lower costs and greater health and functional gains.
Mann et al (2006)	A5	To determine the relationship between fear of falling and neuroticism	Women aged 70 and over	1,691	Non-experimental analytic study (Logistic regression analysis)	A sample of Community dwelling females in the UK	Likert scale (fear of falling) and Eysenck personality inventory (neurotism)	Factors significant in predicting falls include neuroticism, history of falling, experience of fracture, need to use both hands to push up to rise from a chair, poor subjective general health (measured by SF12) and living alone.
Matthews & Brayne (2005)	A5	Providing robust measures of the variation of Dementia across sites in England and Wales	Older people aged 65 years or older	1,463	Non-experimental analytic study (Maximum-likelihood analysis)	Family Health Services Authority lists in five MRC CFA sites in England and Wales (Cambridgeshire; North Wales (Gwynedd); Nottingham; Newcastle and Oxford)	Incidence rate of Dementia, age, sex	There is no variation in dementia incidence across sites within England and Wales. There is no relationship either between Dementia incidence rates and age.
Matthews et al (2004)	A5	Identifying potential biases in two-year follow-up interviews of an aged population	Older people aged 65 or over	12, 988	Non-experimental analytic study (Simple statistical analysis of longitudinal data)	Five diverse centres in England and Wales (East Cambridgeshire, Gwynedd, Newcastle-upon-Tyne, Nottingham and Oxford)	Initial non-response, Longitudinal attrition, age, mini mental state examination, Activities of daily living score, gender, smoking	The sources of bias were cognitive ability, estimates of movements from own home to residential homes, estimates of incidence, longitudinal effects of health and psychiatric diseases.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Mayhew and Lawrence (2006)	A5	An estimation of the changes on provision and costs of intermediate care as a result of reductions in acute hospital care through prevented admission and early discharge	Older people aged 75 and over	A number of workshops	Non-experimental analytic Study (Workshops)	London Borough of Brent with a population of 260,000	Costs of intermediate care packages & acute hospital admission; physical resources	Sound implementation of in intermediate care would result in cost saving to the health economy.
Peet et al (2002)	A5	Understanding the processes, outcomes and costs of a number of alternatives to the eight schemes and identifying elements of best practice of intermediate care in Leicestershire and Rutland	Older people aged 65 and over	Eight intermediate care schemes in Leicestershire and Rutland	Non-experimental analytic study (simple comparative study)	Eight admission avoidance, early discharge and community reablement schemes in Leicester city; Leicestershire and Rutland	Patient's length of stay; Barthel ; source and reasons for referral; extended activities of daily living; discharge destination; EuroQol scores	The majority of intermediate care service users were discharged to their own homes and also had major improvements in their functional and health status. Overall outcomes for all service users were positive. Some problems identified were staff recruitment, inappropriate referrals, few users from ethnic minorities and problems with the physical environment.
Roberts (2002)	A5	To assess user participation in health and social care – getting views of older people on discharge from hospital	People aged 70 years and over	260	Non-experimental analytic study (Simple statistical analysis)	A District General Hospital in the UK	Participation; representation; access; choice; information and redress	Most respondents felt that they had been involved in decisions concerning their discharge.
Tomassini (2006)	A5	A comparison of individual characteristics of the oldest old (85 years and over) to those of the older population as a whole (65 years or over)	Older people aged 85 years or over	2482	Non-experimental analytic study (Logistic regression analysis)	Cross-sectional data from the General Household Survey (GHS)	Visits to the GP, hospital outpatient visits and living arrangements.	Oldest old are more likely to live alone, be widowed, have long-standing illnesses, have higher hospital outpatient visits and fluctuating GP visits.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Tomassini (2005)	A5	A study of the health and other socio demographic characteristics of the oldest old in the UK (i.e. those 85 years or over)	Older people aged 85 years or over	Up to 65,700	Descriptive study (Simple statistical analysis)	Office for National Statistics (2003), Government Actuary's Department (2003), 2001 census, General Register Office for Scotland, Northern Ireland Statistics and Research Agency	Mortality, Health, Geographic distribution.	Oldest old are the fastest growing age group in the UK; more women in private households live alone than men; only 20% live in communal establishments; Most of the oldest old have limiting longstanding illnesses; the highest proportion of this age group is in Southern England and in some Welsh areas.
Wilson et al (2003)	A5	Comparison of the performance of an admission-avoidance hospital-at-home scheme one year after the end of a randomised trial with performance during the trial	Older people with median age at least 81.	78	Non-experimental analytic study	Admission-avoidance hospital-at-home scheme in Leicester, UK.	Baseline characteristics; survival; Barthel index, Sickness Impact Profile 68, Philadelphia Geriatric Morale Scale; Length of stay and visits from a general practitioner	There were no significant difference in the results of the trial and post-trial periods except for higher volume of work and shorter length of stay for the latter.
Griffiths and Wilson-Barnet (2000)	A6	Assessment of what influences length of stay in nursing-led inpatient units (NLIUs)	Older people with mean age above 76 years	296	Non-experimental analytic study - (Analysis of covariance analysis)	NLIU in an acute hospital trust that spread over two sites	Length of stay;	Length of stay appears to be associated with the NLIU's location, staffing levels and patient population.
Matthews et al (2005)	A6	Assessing the association between socioeconomic status and morbidity on older people	Older people aged 75 and over	719	Non-experimental analytic (longitudinal) study	Older people registered with the one large primary care practice in Melton Mowbray, Leicestershire, UK.	Activities of daily living (ADL); age; Place of residence; Housing tenure; Occupation; Income	There was association between disability on one hand and age, housing tenure, living status and income adequacy in multivariate analysis on the other. Income adequacy had the strongest association with disability in longitudinal analysis.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Round et al (2004)	A6	Comparison of patient-based outcomes at six months following emergency admission to a district general hospital or a community hospital	Older people aged 70 and above	376	Non-experimental analytic study - prospective cohort	A district general hospital and five community hospitals in Devon, UK	Euro-QoL EQ-5D; SF-36; Mortality; Place of residence; Number of investigations; Prescribed medications during hospital stay	Quality of life and mortality in community hospitals was similar to those in district general hospitals.
Walker and Jamrozik (2005)	A6	Evaluation of the effectiveness of screening for risk emergencies in the Keep Well At Home (KWAH)	Older people aged 75 and over	2,307	Non-experimental analytic (cohort) study	An intermediate care project comprising 20 of the 38 practices in a West London Primary Care Trust (Hammersmith and Fulham Primary Care Trust)	Levels of coverage; rates of emergency admissions to hospital; attendances at accident and Emergency departments; Emergency admissions of elderly patients	The KWAH project has not been effective in reducing emergency admissions for the elderly.
Walker et al (2005)	A6	Determining efficacy of the Keep Well At Home project considering the pattern of use of Accident & Emergency (A&E) services by those who had been screened by the project.	Older people aged 75 and over	5,373	Non-experimental analytic (cohort) study	An intermediate care project comprising 20 of the 38 practices in a West London Primary Care Trust (Hammersmith and Fulham Primary Care Trust)	Patterns of emergency care	There was a 51% increase in the crude rate of emergency admissions in the first year after screening compared to 12 months before assessment which was not expected.
Young et al (2003)	A6	Estimating the need for post-acute intermediate care in an elderly department for older people	Older people aged over 77 years	4,204	Non-experimental analytic (Prospective) study	An elderly care department in North Bradford Primary Care Trust, UK	Post-acute needs (Rehabilitation, new care home, palliative care, respite/convalescence care)	Only 312 individuals (25.8% of all acute admissions) had post-acute needs. The greatest of these needs was rehabilitation.
O'shea (2004)	B4	A systematic examination and assessment of the 'Senior help line' in Ireland	Older people over 65 years	110	Descriptive study (Simple statistical analysis)	Volunteers registered with Senior Help Line in Ireland	Costs; indicators of best practice	The senior help line has significantly contributed to the health and wellbeing at relatively low cost.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Elkan et al (2001)	C2	Evaluation of the effectiveness of home visiting programmes	Older people aged 65 years or more	15 Studies	Systematic review of trials (randomised and non-randomised) (Meta-analysis)	Medline; Cinahl ; Embase (1980-97); Cochrane Library; Internet; hand search in 'Health Visitor' journal; Reference lists of review articles, key individuals and organisations (for unpublished work)	Mortality, admission to institutional care , hospital, residential nursing homes; functional status, health status; Functional ability;	Older people's mortality and admission to long term institutional care can be reduced by home visits.
Parker et al (2000)	C2	Evaluation of costs, quality and effectiveness of different locations of acute, subacute and postacute care and rehabilitation of older people	Older people aged 65 years or over	84 papers	Systematic review of randomised and pseudo-randomised studies (Mantel-Haenszel method)	MEDLINE; Embase; Cinahl,; SSCI; Psychlit; Other database; 1998/99 searches; Reference lists of papers; hand searching; other including reviewers	Mortality, destinalional outcome, re-admission and costs to the health service.	There is weak evidence on effectiveness, costs for longer standing models. There is also need for systematic review techniques in the area of service delivery and organisation.
Wilson-Barnett et al (2001)	C2	A review of the results from three studies that evaluated NLIUs in south and north London, UK	Older people aged 65 and over	3 studies	A literature review of three studies (controlled trials) that evaluated NLIUs in the UK	Griffiths and Evans, 1995; Griffiths et al. 2000; Griffith and Wilson-Barnet 1998	Length of stay; number of complications; independence; number of discharges; costs;	The first study (Griffiths and Evans, 1995) showed that NLIUs has shorter stay, fewer complications; higher levels of independence and fewer discharged to nursing homes The last two studies (Griffiths et al. 2000; Griffith and Wilson-Barnet 1998) showed that compared to usual care wards, NLIUs were associated with a longer stay and slightly lower daily costs.

Table 3.2: Studies selected after Stage IV

Study	Final Class	Primary Focus	Population	Sample size	Study Design (Key Analysis)	Data Sources	Key Outcomes	Key Findings
Campbell et al (2004)	C5	Investigation of factors that have a significant influence on outcomes of older medical patients admitted to hospital	Patients aged 65 and over who have recently been admitted to hospital	313 papers	Systematic Literature Review Of Prospective Cohort studies	Medline 1966-2000; Cinahl 1982-2000; Web of Science 1081-2000; Reference lists of relevant papers, hand search of 'Age & Ageing' 1074-2000	Function, cognition, depression, illness severity, nutrition, social elements, aspects of diagnosis and demographic details	In addition to routinely available statistics like age, gender and diagnosis, functional Status and cognitive function are important factors that affect outcomes of older medical patients admitted to hospital.
Oliver et al (2004)	C5	To determine what papers have published on risk factors and risk assessment tools for falls in hospital inpatients as well as those on clinical risk assessment tools or individual clinical risk factors that predict falls	Age range not given for all	13 papers	Systematic Literature Review – Cochrane methodology (Simple statistical analysis)	MEDLINE, Embase, Cinahl databases (1966-2002), Cochrane library and Science Citation Index.	Risk factors, Risk assessment tools,	Very few significant risk factors of falls were identified. Predictabilities with sensitivities and specificities of over 70% were shown for simple risk assessment tools.

Table 3.3: Statistical Problems Identified

Study	Statistical Issue									
	Accuracy/ Unreliability of (cost) data	Proxy Issues	Distribution of (Cost) variable	Lack of Generalisability	Bias			Causality	Sample Size/ Lack of power	Statistical Issue Highlighted
					Selection bias	Attrition / Non response bias / Missing Data	Other			
Beaumont and Kenealy (2004)						✓			✓	Relatively small sample; Missing data; Attrition
Beech et al (2004)	✓					✓				Missing data; unreliable data sources
Boston et al (2001)				✓						Lack of Generalisability
Breeze et al (2004)						✓	✓	✓	✓	Small cells for some variables; Missing information; Reverse causation; bias
Carpenter et al (2003)						✓				Incomplete data
Chen et al (2005)	✓						✓			Missing data; Conservative estimates
Cotton et al (2000)			✓			✓				Non-normality; Attrition
Downing and Wilson (2005)						✓	✓			Missing data; Differences in size of samples and other differentiation issues (sample differentiation?)
Elkan et al (2001)							✓			Heterogeneity
Fleming et al (2004)				✓						Generalisability
Fletcher et al (2004)	✓					✓	✓			Lack of national data; no control group; outcomes irrelevant in certain circumstances; underestimation of hospital and institutional admissions; interpretation of QoL measures

Table 3.3: Statistical Problems Identified

Study	Statistical Issue									
	Accuracy/ Unreliability of (cost) data	Proxy Issues	Distribution of (Cost) variable	Lack of Generalisability	Bias			Causality	Sample Size/ Lack of power	Statistical Issue Highlighted
					Selection bias	Attrition / Non response bias / Missing Data	Other			
Forder and Netten (2000)	✓	✓	✓	✓			✓			Bias; Skewness of price distributions; Heteroscedasticity; Proxies; Unreliable data; Sample Representativeness
Green et al (2005)			✓	✓		✓	✓			Non-normality; Attrition; Missing data; Lack of generalisability; Other bias
Griffiths et al (2000)			✓			✓				Non-normality; Attrition
Griffiths et al (2001)	✓		✓							Uncertainties in costs estimates; non-normality;
Griffiths and Wilson-Barnet (2000)			✓							Distribution of length of stay variable
Gunnel et al (2000)			✓						✓	Non-normality; Sample size
Grundy and Sloggett (2003)	✓			✓	✓	✓		✓		Missing data; Unrepresentative samples; Selection bias ; unreliability due to self-reporting; difficulty in establishing causal pathways between variables
Harris et al (2005)	✓	✓								Different methods of calculating costs (bottom-up, top-down and bottom-up plus discounting); Proxies
Kaambwa et al (2008)	✓		✓			✓				Missing data; skewness; Unreliable cost data; Heteroscedasticity

Table 3.3: Statistical Problems Identified

	Statistical Issue									
Study	Accuracy/ Unreliability of (cost) data	Proxy Issues	Distribution of (Cost) variable	Lack of Generalisability	Bias			Causality	Sample Size/ Lack of power	Statistical Issue Highlighted
					Selection bias	Attrition / Non response bias / Missing Data	Other			
Mann et al (2006)	✓			✓		✓	✓	✓		Causation cannot be inferred from a cross-sectional study; Participant and response bias; unreliability of self-report data; generalisability
Matthews and Brayne (2005)		✓				✓			✓	Small sample size, Use of proxy informants; sample attrition
Matthews et al (2004)		✓				✓	✓			Attrition bias; non-response bias; multiplicity/choice of outcome measure to use; use of proxy interviewees
Matthews et al (2005)				✓		✓		✓		Attrition; Missing data; Unrepresentative sample
Miller et al (2005)			✓	✓						Skewness of cost data; Results not generalisable
Newton et al (2006)	✓		✓				✓			Distributional issues (Skewness of rate of fallers and costs); Data collection period not homogenous to that where data were not collected; Sample is Heterogeneous to those for which data has not been collected; underestimation of outcomes because of un-accounted for interventions
O'Reilly et al (2005)	✓			✓						Uncertainty or unreliability of cost and other variables; Lack of generalisability
O'Reilly et al (2008)	✓					✓				Unreliable cost estimates; missing data

Table 3.3: Statistical Problems Identified

Study	Statistical Issue									
	Accuracy/ Unreliability of (cost) data	Proxy Issues	Distribution of (Cost) variable	Lack of Generalisability	Bias			Causality	Sample Size/ Lack of power	Statistical Issue Highlighted
					Selection bias	Attrition / Non response bias / Missing Data	Other			
Parker et al (2000)				✓		✓			✓	Missing data; small sample size; Varied outcomes; Generalisability of results
Peet et al (2002)						✓				Missing data
Rait et al (2005)		✓				✓	✓			Use of cut-offs; missing data; Use of self-reported data; use of other (proxy) instruments; non- completion
Richardson et al (2001)	✓		✓			✓			✓	Non-response; Uncertainty of cost estimates; Positively skewed costs; Missing data; Lack of power; Heteroscedasticity
Roberts (2002)				✓		✓	✓			Not able to limit sampling process; missing data; representativeness
Round et al (2004)			✓	✓		✓				Non-normality; Skewness; Incomparability; Attrition; missing data
Seshamani and Gray (2004a)						✓	✓			Right censoring; Attrition; missing variables in model (model specification)
Seshamani and Gray (2004b)			✓			✓				Missing data; Skewness of cost data; Multicollinearity; correlations
Shaw et al (2000)			✓	✓			✓			Non-normality; Lack of generalisability; Blinding not feasible

Table 3.3: Statistical Problems Identified

	Statistical Issue									
Study	Accuracy/ Unreliability of (cost) data	Proxy Issues	Distribution of (Cost) variable	Lack of Generalisability	Bias			Causality	Sample Size/ Lack of power	Statistical Issue Highlighted
					Selection bias	Attrition / Non response bias / Missing Data	Other			
Tomassini (2006)	✓	✓		✓		✓	✓		✓	Inappropriate data; inadequate numbers in survey; missing data; Bias; Low response rates; Use of proxies; representativeness of results
Wade et al (2003)						✓			✓	Attrition; Reduced power
Walker and Jamrozik (2005)	✓					✓				Missing data; Inaccurate data
Walker et al (2005)						✓				Non-response; Missing data;
Walsh et al (2005)	✓								✓	Lack of detailed cost breakdown; use of top-down cost estimates; sample size
Wilson et al (2003)						✓				Missing data
Young et al (2003)				✓						Lack of generalisability
Young et al (2007)			✓				✓		✓	Small sample size; Outliers; Other bias

CHAPTER FOUR - CASE STUDY: A NATIONAL EVALUATION OF THE COSTS AND OUTCOMES OF INTERMEDIATE CARE SERVICES FOR OLDER PEOPLE – DESCRIPTION OF STUDY

4.1 Introduction

This chapter describes the National Evaluation of the Costs and Outcomes of Intermediate Care Services for Older People in the UK (ICNET 2005) which has provided the data analysed in chapters five, seven and nine. An appreciation of the source of these data will be helpful in understanding and interpreting the results presented in these three chapters. The rationale and objectives for this evaluation are summarised and the research activities that were carried out to meet the objectives are outlined. Descriptive statistics from the evaluation are also presented. The chapter also discusses the statistical problems that were identified in the demonstration dataset, which problems were the subject of the empirical analyses, before closing with the key messages to health services researchers working with similar datasets.

4.2 Rationale

As described in section 2.5 of chapter two, intermediate care can be viewed as a community or institution-based service that represents an alternative to high cost in-patient hospital care. As such, one would expect that compared to hospital care, appropriate use of intermediate care would almost certainly be associated with lower expenditures and higher effectiveness. But as the literature review of evaluations of intermediate care services in the UK also reported in chapter two revealed, evidence on the performance of these services relative to alternatives is mixed. As a result, studies such as Parker et al (2000) and Grimely and Tallis (2001) called for more evaluations to be conducted so as to provide more data on the effectiveness and cost effectiveness of intermediate care services.

4.3 Objectives

It was against this background that the national evaluation of the costs and outcomes of intermediate care services for older people in the UK was born. There was a need to fill the gap that existed in the literature concerning the effectiveness and cost effectiveness of intermediate care. As part of the Health Policy Research Programme, the Department of Health and the Medical Research Council commissioned a series of research projects to evaluate intermediate care. Upon successfully bidding for one of the projects, a team made up of researchers from the Universities of Birmingham (Health Services Management Centre - HSMC) and Leicester (Nuffield Community Care Studies Unit) conducted a national evaluation. Two other research projects were commissioned at the same time. One was a comparative case study and national audit of intermediate care expenditure which was led by Professor Gerald Wistow from the Nuffield Institute for Health at the University of Leeds. This project took a 'whole systems' approach and focussed on the structure, process, outcomes and cost effectiveness of intermediate care for older people (Department of Health 2002b). The spotlight was on the impact of service system, service components and individual patient/user and caregiver on intermediate care. In addition, the project sought to adopt a comparative case design and also conduct a national audit of intermediate care expenditure so as to meet its objectives (Department of Health 2002b). The other project looked at the effectiveness of community hospitals in providing intermediate care for older people and used a mixed-methods study approach. This project was led by Professor John Young from St Luke's Hospital, Bradford. The aim of this evaluation was to provide evidence of the extent to which community hospital care can foster independence and lead to a reduction in institutionalisation of older people (Department of Health 2002b).

The research work for the national evaluation of costs and outcomes of intermediate care was carried out between January 2002 and December 2004 and had five major objectives which can be divided into qualitative and quantitative components.

The objectives of the qualitative part were to:

- (i) establish the range, spread and speed of development of intermediate care services across England,
- (ii) explore the views of intermediate care leads on the benefits and challenges of implementing intermediate care policy and
- (iii) assess the impact of intermediate care on the service system as a whole and on individual service users.

The quantitative part of the study sought to:

- (iv) explore the costs of intermediate care schemes in relation to their outcomes and
- (v) synthesise evidence from this and other research on the costs and outcomes of different models of intermediate care and on best practice.

This author was involved in research that met objectives (iv) and (v).

4.4 Methods

After approval by the Trent Multi-Centre Research Ethics Committee, three main research activities were pursued in order to meet the objectives described above:

- Postal survey of intermediate care co-ordinators (objectives i and ii).
- Case studies with ‘whole systems’ (Primary Care Trusts - PCTs) of intermediate care (objectives iii and iv).
- Systematic review of evidence for the effectiveness of intermediate care (objective v).

The postal survey of intermediate care co-ordinators was planned for the early stages of the project and was meant to guide the choice of case-study sites. A simple and short questionnaire comprising a combination of both fixed-choice and open response questions was prepared. This was to allow for both quantitative and qualitative data to be collected. Telephone calls were made to every PCT in order to identify informants and in the end, 106 questionnaires were completed. This represented a 46% response rate.

Case-study research in ‘whole systems’ (PCTs) of intermediate care involved studying areas with a specific geographical boundary in some depth and was performed so as to gain a better understanding of each area’s effect in terms of implementation, outcomes and costs. The following criteria were used to choose the PCTs (case-study sites) that would participate in the evaluation:

- A range of intermediate care services operational for at least 2-3 years.
- Reasonable throughput into the intermediate care system (at least 1000 cases per annum).
- A mix of urban and rural sites.

- Senior management support for the collection of routine data by services themselves.
- Clinical and managerial support for participation in the national evaluation.

A summary of the case study sites is presented in Table 4.1.

A systematic review was conducted and updated the work of some members of the research team who had previously published a systematic literature review on the “Best Place of Care” for older people after acute illness and with subacute illness (Parker et al. 2000). In this review, however, the focus was on UK studies and on intermediate care of patients which had developed since the “Best Place of Care” review was conducted in 1999.

4.5 Data collected

Data were obtained from five case study sites in which the qualitative study (Regen et al. 2008) was also undertaken. In order to provide evidence on the costs and outcomes associated with intermediate care, case studies of ‘whole systems’ of intermediate care were used. Data were collected on an observational cohort of consecutive admissions to intermediate care services within the case studies. Data on resource use and costs, and outcomes (measured using a generic quality of life instrument - the EQ-5D, and a general measure of functioning - the Barthel Index), were collected from a sample of 2253 episodes of intermediate care in five case study evaluation anonymous sites. Data on patient characteristics, descriptors of intermediate care service as well as intermediate care-related services’ variables were collected using a Proforma (ICNET 2005). Unit costs were obtained from service budgets and from national

sources. Staff working for the intermediate care services collected the quantitative data according to protocols set out by the evaluation team. For all intermediate care admissions over a defined period, service staff completed a study proforma, with or on behalf of their patients, at the point of admission to the service. They completed discharge questions on the day of discharge, transfer or as soon as possible following end of service provision. Data collection in all five sites continued for approximately seven months.

Data were collected on:

- the likely service provision were the intermediate care service not available (as a check on admission appropriateness);
- patient characteristics (age, sex, living arrangements, and health at admission and discharge);
- descriptors of the intermediate care service (admission avoidance or supported discharge, and residential or non-residential);
- features of the intermediate care episode (duration of service provision, transfer before end of intermediate care episode, patient death, and episode completed);
- source of referral (primary care, secondary care or social services); and.
- support services received at the start of an intermediate care episode (home care, district nurse visit, domiciliary therapy, meals on wheels and others).

Because of non-response or complete lack of information, some missing data were encountered. Ways of dealing with this issue are the subject of chapters six and seven.

A full list of all the variables on which data were collected together with their level of completeness is presented in Table 4.2.

While many other measures of outcomes could have been used for populations of older people such as the general health questionnaire (Grundy & Sloggett 2003), Nottingham extended activities of daily living scale (Fleming et al. 2004), sickness impact profile (Wilson et al 2003) and Nottingham health profile (Green et al 2005), the two outcome measures that were selected were the Barthel Index and the EuroQol EQ-5D. The former is a functional assessment of activities of daily living instrument that has also been widely used to measure health outcomes in elderly populations (Yohannes et al. 1998; van den Bos & Triemstra 1999; Davies 1996 and Lyons et al. 1997). The latter is a generic quality of life instrument which has been used previously to model health-related quality of life for older people (Coast et al. 1998a; Brazier et al. 1996; van den Bos & Triemstra 1999; Lyons et al. 1997). Most clients were incapable of self-completing the EQ-5D questions and therefore stated their responses to staff that in turn filled out the questionnaires. The Barthel index questionnaire was filled out by staff.

The Barthel index has been recommended for scientific research purposes and also for use as a proxy for other outcome measures (Wilkinson et al. 1997). To measure a person's level of functional independence, the Barthel uses 10 items, with each item carrying different weights (Mahoney & Barthel 1965). These items are feeding, bathing, grooming, dressing, bowels, bladder, toilet use, transfers (bed to chair and back), mobility (on level surfaces) and stairs. Two items are rated on a two-point scale of 0 and 5, six on a three-point scale of 0, 5 and 10 and the last two items are rated on

a four-point scale of 0, 5, 10 and 15. The scores/scales on each item are added to produce an overall score which ranges from 0 to 100. To standardise them, the scales used in this study were divided by 5 and therefore ranged from 0 to 20. The higher the score recorded for an item, the greater the level of independence. The reliability, sensitivity and suitability for proxy-assessment of the Barthel has been shown elsewhere (Mahoney & Barthel 1965; Shah et al. 1989; Wolfe et al. 1991). A copy of the Barthel index questionnaire is presented in A3 of the appendix.

The EQ-5D is comprised of five dimensions of health: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. There are three levels of impairment in each domain: no, some/moderate, and extreme problems in the relevant dimension of health. Using these responses, the EQ-5D is able to distinguish between 243 states of health (Dolan 1997; Brazier et al. 2004). The UK-specific EQ-5D valuation algorithm was used in order to convert the EQ-5D health description into a valuation (Dolan 1997). After normalisation, EQ-5D scores have a range of -0.59 to 1: the maximum score of 1 represents perfect health and a score of 0 represents death. Scores less than 0 represent health states that are worse than death (Dolan 1997; Kind et al. 1999; Murphy et al. 2001; Post et al. 2001). Its generic nature makes it comparable across patient populations. A copy of the EQ-5D questionnaire is presented in A4 of the appendix.

Health service costs were estimated for every patient episode. These estimates were calculated using data on patient-specific resource use, as collected by proforma, multiplied by appropriate unit costs. The latter were obtained from intermediate care service-specific top-down unit cost estimates from budget statements for the

individual intermediate care services, and from published national sources (Curtis & Netten 2004). A common price year of 2004 was used for all unit costs.

4.6 Baseline Patient Characteristics

Baseline demographic and other details of patients are presented in Table 4.3. Across all five case study sites, data were collected on 2253 episodes of intermediate care. The majority of episodes (82%) were non-residential, and approximately 55% of episodes were for admission avoidance services and 42% were for supported early discharge services³ with the remaining 3% receiving other intermediate care services such as rehabilitation. In terms of the relationship between setting and function, non-residential settings accounted for 89% and 72% of all admission avoidance and supported discharge services, respectively. The median age at the start of an intermediate care episode for the total sample was 81 years, with three-quarters aged 75 and over. Those in services providing an admission avoidance function tended to be slightly older ($\chi^2 = 39.9$, $p < 0.01$). About 70% of all admissions were for female patients, and approximately half of the sample lived alone at the time of start of an intermediate care episode. The use of support services was, unsurprisingly, higher among patients admitted to admission avoidance schemes, given the community setting for such schemes ($\chi^2 = 82.1$, $p < 0.01$). The mean EQ-5D and Barthel scores for the total sample were 0.42 and 14.8, respectively, indicating that on admission patients were in a relatively poor state of health and/or had poor levels of functioning. Acute admission avoidance patients tended to have poorer health status on admission ($\chi^2 = 67.3$, $p < 0.01$) and were more dependent than supported discharge patients ($\chi^2 = 47.1$, $p < 0.01$).

³ Admission avoidance and supported discharge services have been defined in chapter one.

4.7 Costs

Table 4.4 reports summary measures of the cost of intermediate care episodes. These data are reported both for the total sample and separately for patients admitted to admission avoidance and supported discharge intermediate care services.

4.8 ‘Appropriateness’ and intermediate care episodes

In terms of appropriateness of the intermediate care admission, defined in terms of whether the alternative would have been hospital-based care, the study data suggest that for 974 patients (i.e. 53% of all patients for whom data on an alternative place of care was given) the use of intermediate care was appropriate. Thus, these data suggest a large proportion of patients (i.e. 47%) in this study were inappropriately admitted to intermediate care services. Table 4.5 reports results split by appropriate and inappropriate intermediate care admissions. In general, it appears that appropriate admissions had a higher intermediate care cost ($\chi^2 = 39.3$, $p < 0.01$). However, this should not be a concern because the appropriate intermediate care admissions are those where patients would otherwise be occupying a hospital bed and so a significant cost is being avoided. They also had a shorter duration of service provision ($\chi^2 = 6.5$, $p < 0.01$) while the Barthel scores were lower at admission than those for inappropriate admissions ($\chi^2 = 12.6$, $p < 0.01$). The latter may explain why the change in functional independence from admission to discharge (as measured by the change in Barthel scores) was significantly higher for appropriate admissions. There does not seem to be any significant difference in the quality of life between appropriate and inappropriate admissions ($\chi^2 = 1.6$, $p = 0.21$).

Further, the appropriate admissions appear to achieve slightly larger gains in terms of Barthel change ($\chi^2 = 18.9$, $p < 0.01$). This pattern of differences between appropriate and inappropriate intermediate care admissions is similar for both admission avoidance and supported discharge forms of intermediate care. The interpretation is that the health gains are larger where the admission is appropriate.

4.9 Outcomes

Table 4.4 also reports summary measures of duration of intermediate care episodes and changes in EQ-5D and Barthel scores. These data are again reported both for the total sample and separately for patients admitted to admission avoidance and supported discharge intermediate care services. The mean duration of intermediate care service provision was 29.5 days. Compared to those in supported discharge schemes, patients in admission avoidance schemes were associated with lower quality of life (mean EQ-5D score of 0.32 vs. 0.50; $\chi^2 = 67.3$, $p < 0.01$) and lower functional status (mean Barthel scores of 13.8 vs., 15.6; $\chi^2 = 47.1$, $p < 0.01$). Further, patients in acute admission avoidance schemes also registered a higher change in EQ-5D (0.23 vs. 0.12; $\chi^2 = 12.5$, $p < 0.01$) and a non-significant higher change in the Barthel index (2.11 vs. 1.42; $\chi^2 = 3.7$, $p = 0.16$).

4.10 Identifying statistical problems present in the demonstration dataset

The national evaluation of costs and outcomes for older people in the UK sought to understand what factors caused variation in costs and outcomes of this population. A regression model framework was chosen as the main form of analysis. It was therefore important to understand what data was available to pursue this work and if

there were any statistical issues that needed to be taken care of to ensure that unbiased results were obtained.

4.10.1 Missing data

As shown in Table 4.2, the final dataset revealed that there was a substantial amount of data missing. In particular, up to 42% and 38% of the data on EQ-5D and Barthel scores, respectively, were missing. Further, data on costs per patient were also missing for 31% of the sample. As can be further seen in Table 4.2, all but one variable (Type of IC) had missing data ranging from 3 to 18%.

4.10.2 Lack of generalisability

Data used was provided by five case studies which had a mix of rural, semi-urban and urban sites (table 4.1). A lot more sites could have been included to make the sample even more representative of the national population. Though the level of generalisability could have potentially been increased, the ICNET dataset represents the largest evaluation of intermediate care done and published in the UK to date. Therefore the problem of lack of generalisability was not of serious concern in this dataset as it was deemed to be the best that was available at the time of analysis.

4.10.3 Unreliability or inaccuracy of estimates

The national evaluation sought to use the bottom-up approach to determine the costs per patient as this would have ensured that the inter-patient variability that is usually inherent in such costs was captured. However, it was not possible to do this due to circumstance beyond the national evaluation team. Therefore, the cost variable was calculated using a top-down approach which method almost certainly lead to under-

representation of the true variability in costs within and between intermediate care services. No other data on costs for the intermediate care services that participated in the national evaluation were available and therefore it was difficult to conduct valid sensitivity analysis. Other inaccuracies in the data were dealt with during the data cleaning stage of the project.

4.10.4 Problems associated with distributional characteristics of variables

4.10.4.1 Ceiling and floor effects

The Barthel and EQ5D scores have minimum and maximum permitted scores. The EQ-5D is an instrument whose values range from -0.59 to 1 while the Barthel score has a scale of 0 to 20. The differences in the scores between admission and discharge therefore range from -1.59 to 1.59 and -20 to 20, for the EQ-5D and Barthel Index, respectively. This implies that there was a possibility of floor and ceiling effects which needs to be taken into consideration when dealing with these outcomes. After examining the data, these effects were evident. In particular, 4% and 7% of the differences in EQ-5D scores were equal to -1.59 and 1.59, respectively. For the Barthel score, 2% and 4% of the scores were equal to -20 and 20, respectively.

4.10.4.2 Too many 'zero' values in the dependent variable

Although there were 355 and 109 observations with values of zero in the EQ-5D and Barthel index outcome data, respectively, this was not a problem as may be the case when dealing with cost or expenditure data where the interest is only in positive values (Mullahy 1998). This is because the value of zero was also of interest since a 'zero' result shows that there was no improvement in either the quality of life or

functional independence. There were no zero costs encountered in the cost per patient variable.

4.10.4.3 Skewness

Line graphs of the kernel density estimates were produced for the two outcome measures and these are shown below (figures 4.1 and 4.3). These show skewness (skewness values $\neq 0$) which results can be observed further from the standardized normal probability (P-P) plots (figures 4.2 and 4.4). Tests of skewness (the Shappiro-Francia test) tested the null hypotheses that the outcome variables were normally distributed. The null hypotheses were rejected for both outcome measures (Δ EQ-5D: z-score = 7.4, $p < 0.001$ and Δ Barthel: z-score = 6.40, $p < 0.001$) confirming that the two dependent variables were non-normal in their distribution.

Figure 4.1: Change in EQ-5D: Graph of kernel density estimates

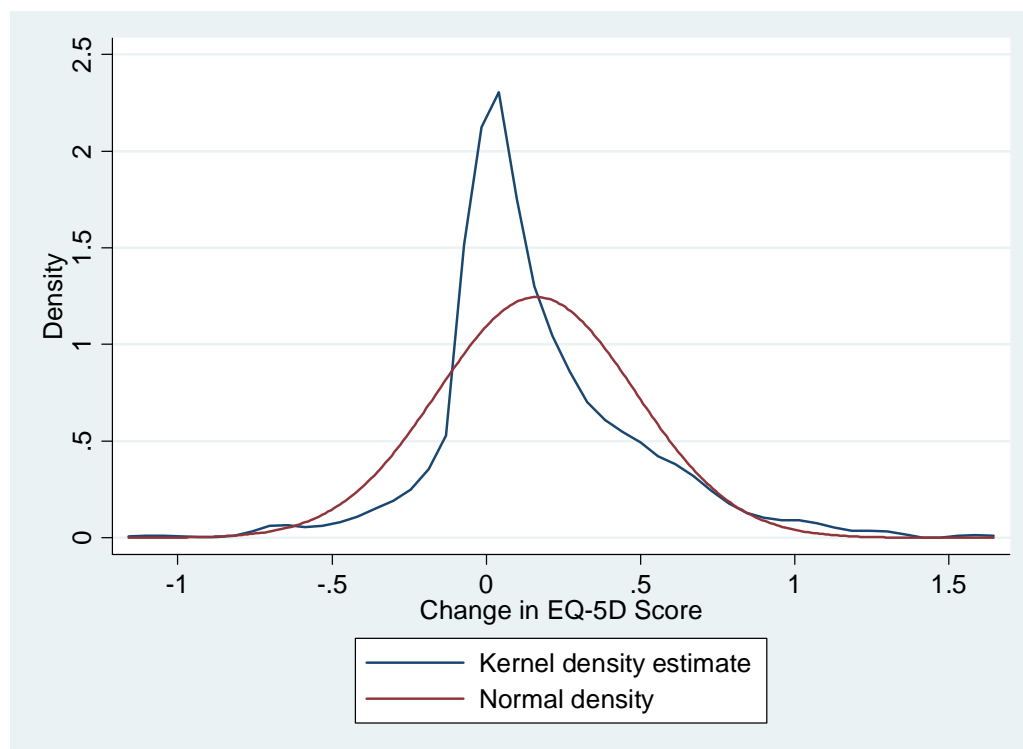


Figure 4.2: Change in EQ-5D: P-P plot

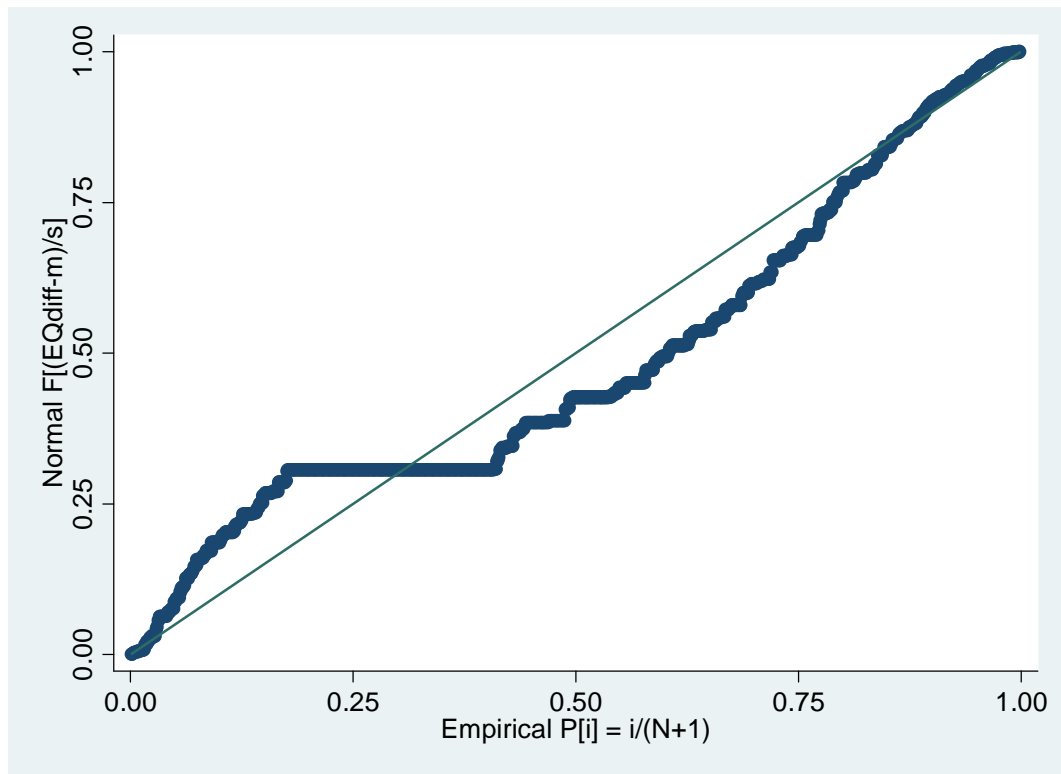


Figure 4.3: Change in Barthel: Graph of kernel density estimates

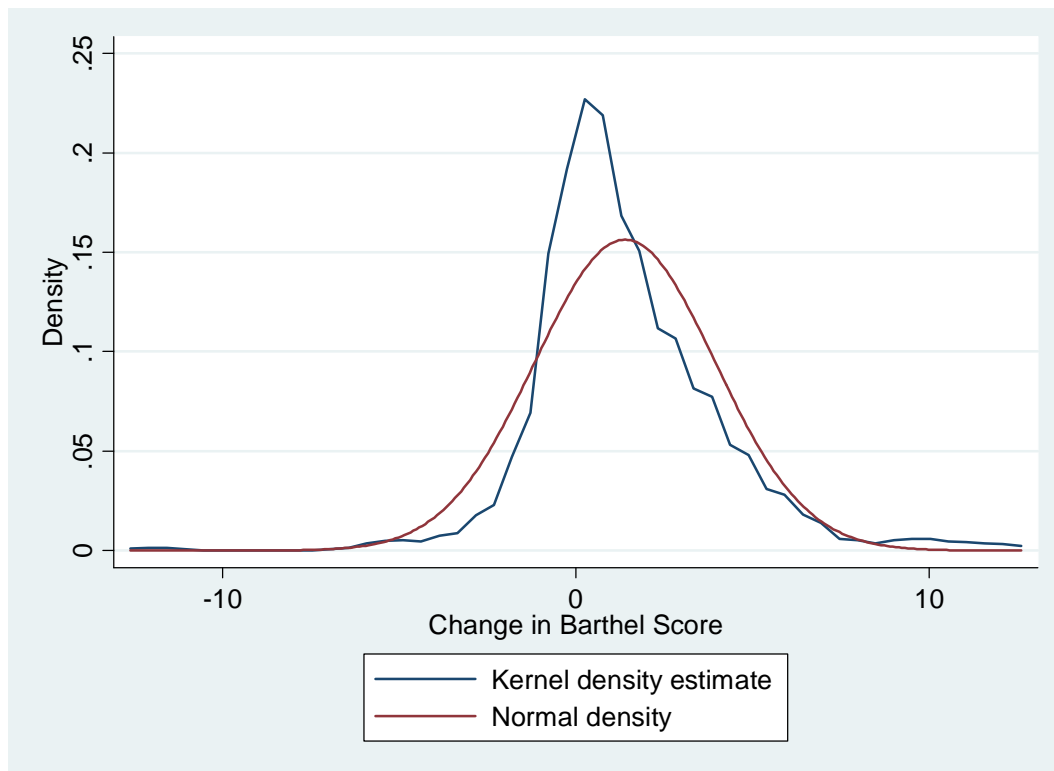
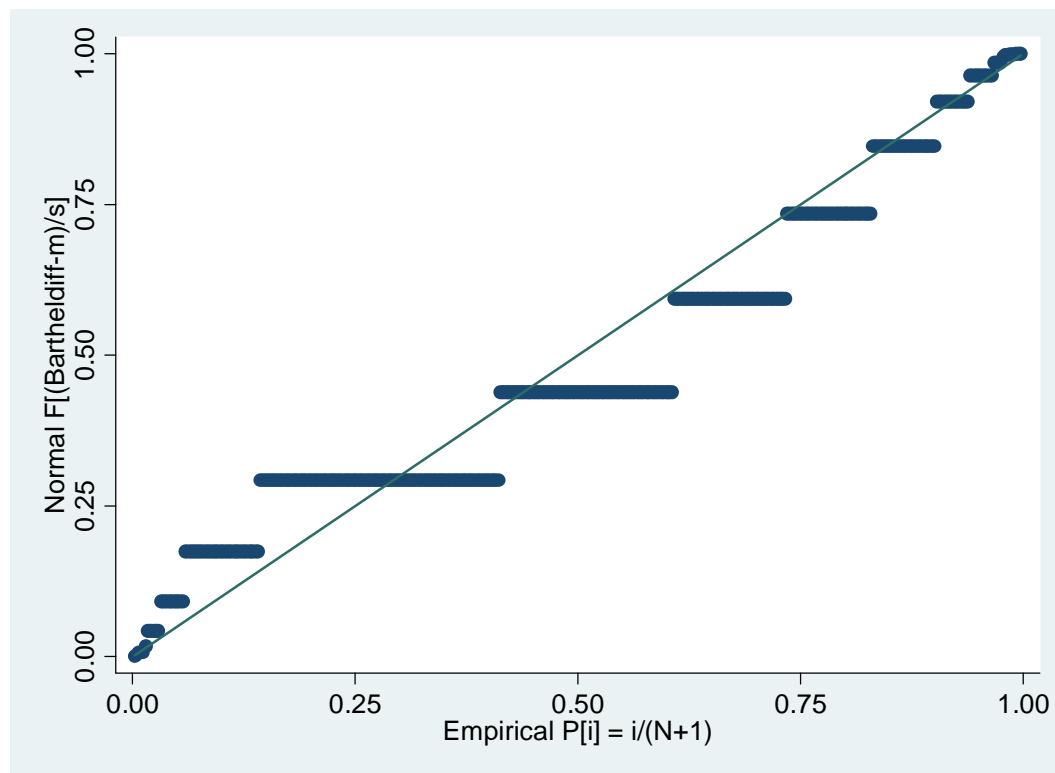


Figure 4.4: Change in Barthel: P-P plot



As shown in figure 4.5, the distribution of costs is right skewed and this is expected for cost data (Thompson & Barber, 2000). This is because the majority of patients had low costs while a few had high costs. Tests of skewness (the Shappiro-Francia test) tested the null hypotheses that the cost per patient variable was normally distributed. The null hypothesis was rejected (z-score = 9.19, $p < 0.001$) confirming that the variable was skewed in its distribution. This result was also born out by the P-P plot (figure 4.6). In terms of costs, it is clear that intermediate care episodes consume considerable resources, with the mean episode cost of approximately £1200 (table 4.4). The difference in cost between admission avoidance and supported discharge schemes is significant ($\chi^2 = 110.6$, $p < 0.01$), with supported discharge cases incurring a higher mean cost in excess of £1500 per episode (table 4.4).

Figure 4.5: Distribution of costs per patient

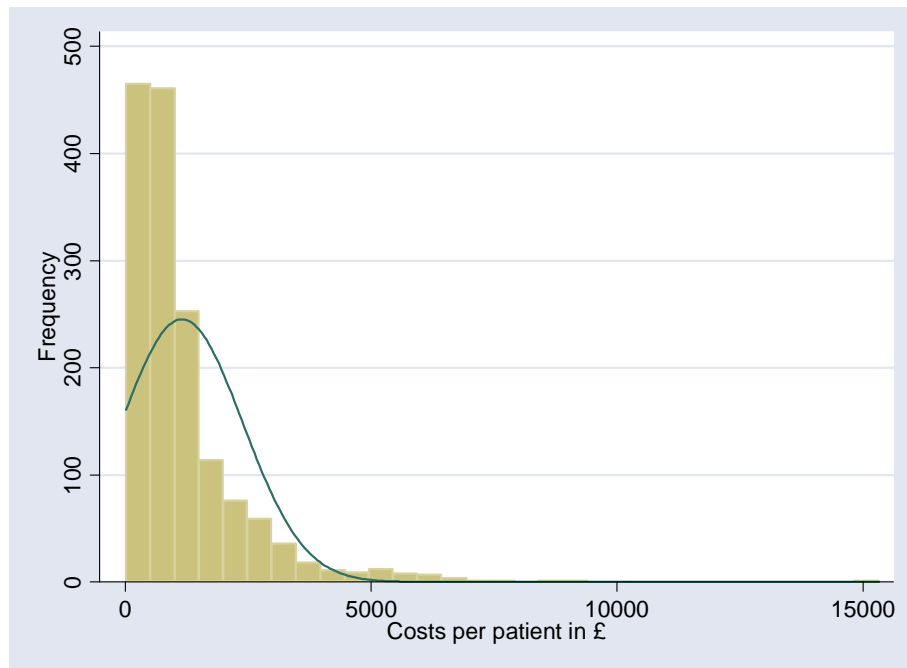
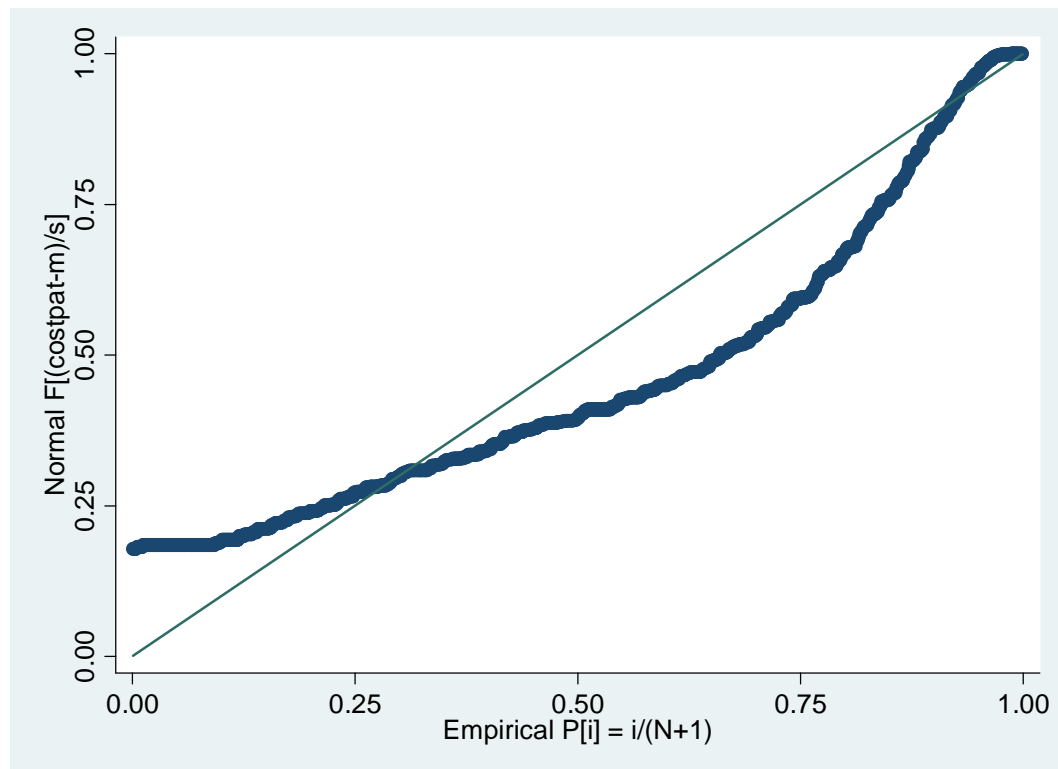


Figure 4.6: Cost per patient: P-P plot



4.10.4.4 Heteroscedasticity

Heteroscedasticity is when the variance of the residuals is non-constant (Gujarati 1995). There are graphical and non-graphical methods of determining whether heteroscedasticity is present or not. Both methods require that one runs a regression model first. A widely used graphical method involves plotting the residuals against fitted (predicted) values of the dependent variable. A pattern to the plots indicates the presence of heteroscedasticity. A number of non-graphical method or tests can be used to detect heteroscedasticity including the Park test, White test, Glejser test or the Breusch-Pagan test (Gujarati 1995; Breusch & Pagan 1979). The plots shown in figures 4.7 to 4.9 all have a hint of some pattern to the plots. In figure 4.7, there is a larger concentration of data points around the residual values of 0 as is the case in figure 4.9 while in figure 4.8, the data points seem to get narrower towards the right end. These patterns suggest the presence of heteroscedasticity. Results from the White test ($\chi^2 = 80$, $p < 0.001$ and $\chi^2 = 112$, $p < 0.001$) confirmed the presence of heteroscedasticity in the EQ-5D and Barthel models, respectively. The White test results for cost per patient also confirmed heteroscedasticity in this variable ($\chi^2 = 431$, $p < 0.001$).

Figure 4.7: Plot of residuals against predicted values – change in EQ-5D

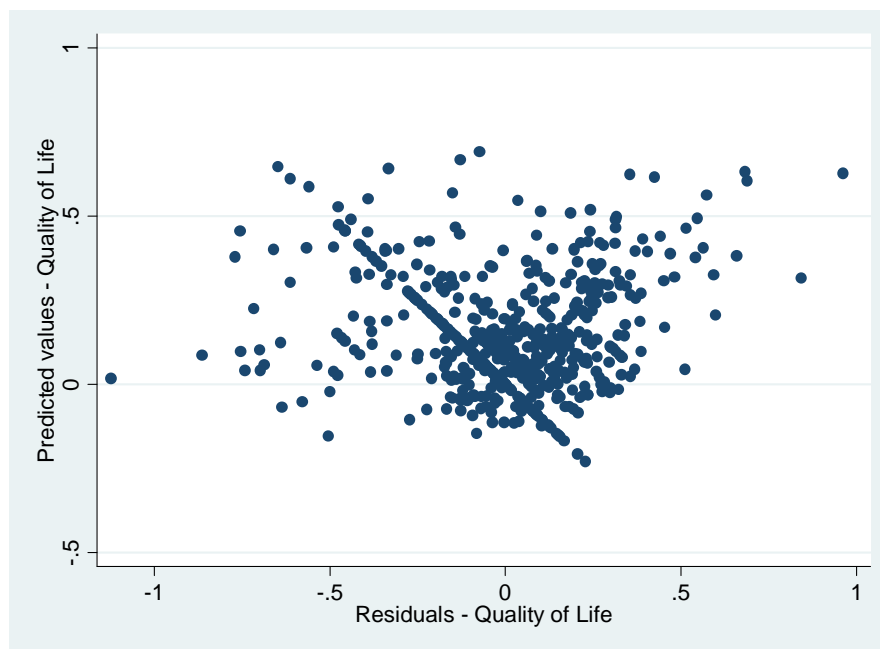


Figure 4.8: Plot of residuals against predicted values – change in Barthel Index

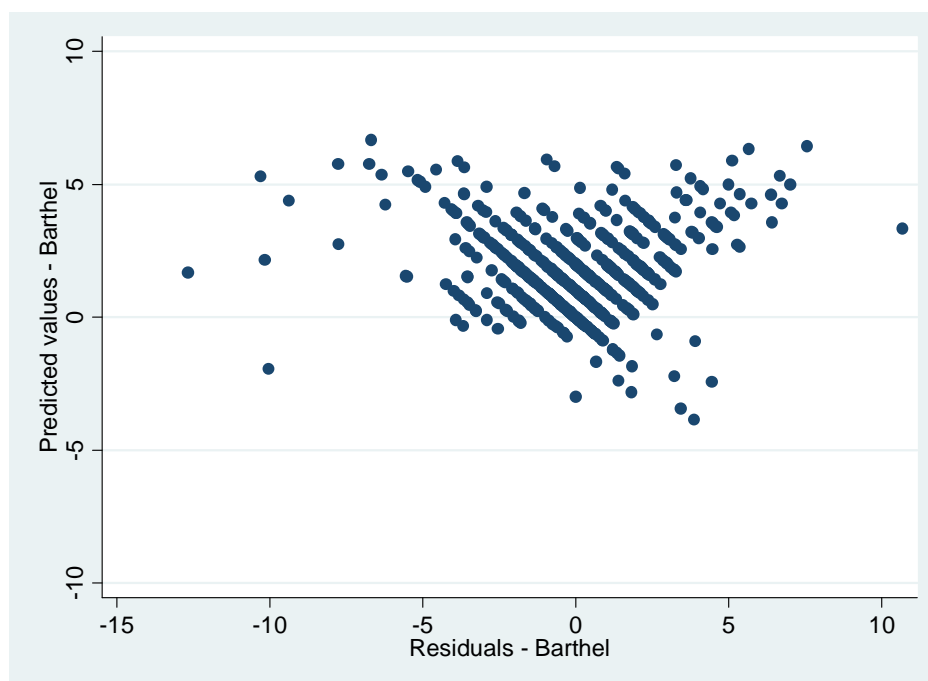
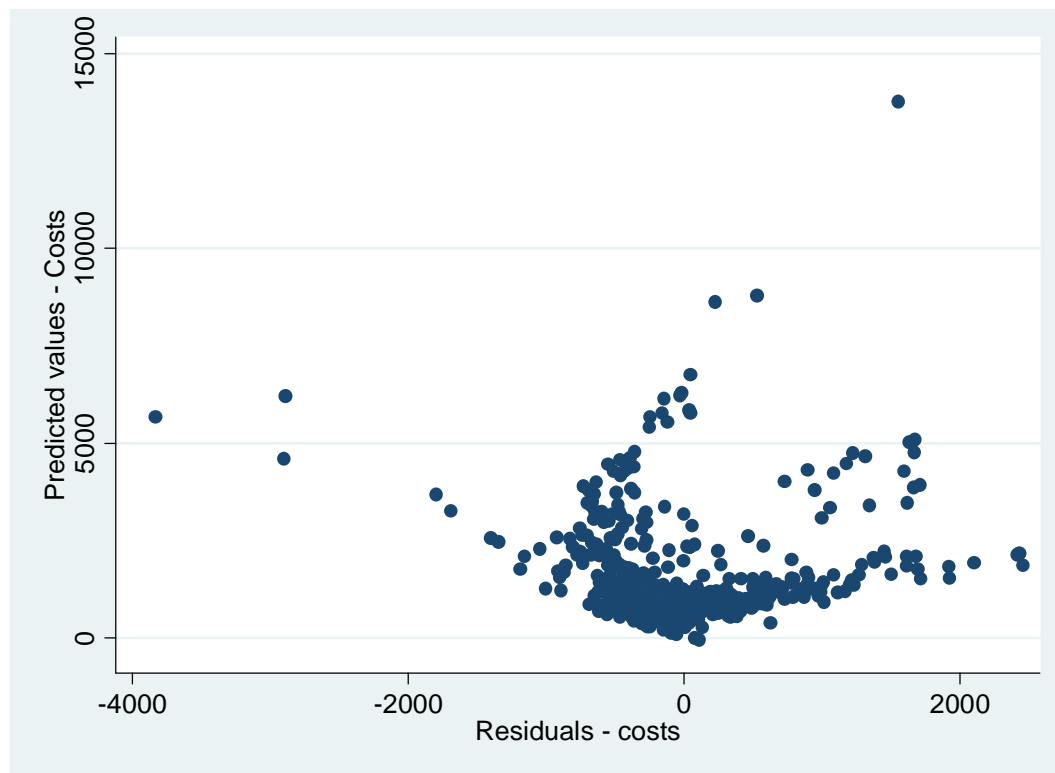


Figure 4.9: Plot of residuals against predicted values – cost per patient



4.10.5 Problems of sample size and lack of power

The sample consisted of 2,253 observations which is a fairly large sample to use in answering the questions that the national evaluation needed to answer. The sheer size of this dataset did not therefore lead to any cause for concern that the observations used were too few or that there would be a problem of lack of power in the analysis.

4.10.6 The need for predicted utility measures of outcome.

Up to 40% of the EQ-5D scores at admission (at the start of an intermediate care episode) were missing. The EQ-5D is desirable for many reasons, most importantly because it can be converted into utilities that can then be used for conducting cost effectiveness analyses. A method of predicting ED-5D scores from other outcome

measures therefore becomes significant. The other outcome measure that was collected in the national evaluation was the Barthel index. Importantly, there was less missingness in the Barthel scores compared to the EQ-5D i.e. only 31% of the Barthel scores were missing at admission. This therefore presented an opportunity to map Barthel scores onto EQ-5D scores.

4.11 Discussion

4.11.1 Key findings

The data from the evaluation revealed that a significant proportion of patients (possibly as high as 47%) being referred for intermediate care in the UK did not meet the stated Department of Health criteria for appropriateness. That is, many patients using intermediate care services would not otherwise be cared for in a hospital-based setting. This finding is supported by the Barthel data for our sample which suggest that patients receiving intermediate care, in routine practice settings in the UK, were less dependent than those seen in earlier trials of hospital at home (Wilson et al. 1999). The median Barthel at the start of an intermediate care episode in our total sample was 16, compared with 9 in a trial of an admission avoidance hospital-at-home scheme (Wilson et al. 1999). This supports a view that intermediate care is not currently being targeted at those most in need. What this analysis revealed was that robust and reliable clinical criteria needed to be developed to ensure appropriate admission of intermediate care patients. In addition, there was need for close co-operation between hospital and community service providers as regards the selection of patients and targeting of intermediate care and acute care services to meet defined clinical need (Kaambwa et al. 2008 in appendix). As opposed to the results of a survey conducted among intermediate care coordinators as part of this evaluation,

these quantitative results suggest that the vast majority of intermediate care services currently being provided are for admission avoidance services. It should be noted however that there is relatively little evidence in the literature on the effectiveness or cost-effectiveness of admission avoidance services.

This chapter also highlighted four statistical problems that were identified in the dataset and are of concern. These were missing data, problems associated with the distributional characteristics of variables, the need for predicted utility measures of outcome and unreliability of the cost per patient variable. The last problem is best addressed using sensitivity analysis. But because no other cost data were available from the case study sites or indeed from comparable sites with which to carry out this sensitivity analysis, a decision was made not to proceed further with addressing this problem in this thesis. Any such attempt is likely to suffer from serious deficiencies. As a result, the rest of the chapters in this thesis consider the first three problems.

4.11.2 Key messages for health services researchers from this chapter

This chapter described the largest evaluation of intermediate care done and published in the UK to date and drew data from five case study sites. Most episodes in the evaluation were non-residential and there were more services performing an admission avoidance function compared to a supported discharge one. The evaluation also revealed that a significant proportion (47%) of intermediate care admissions were inappropriate. Four of the seven major statistical problems identified in the literature review reported in chapter three were also found in the demonstration dataset that was drawn from the evaluation. These were missing data, problems associated with the distributional characteristics of variables, the need for predicted utility measures of

outcome and unreliability of the cost per patient variable. The thesis only considers the first three problem as enough data were not available to tackle the last one.

4.12 Conclusion

This chapter described the case study that provided the data used in the empirical analyses of this thesis. Several descriptive statistics on baseline patient characteristics, costs, outcomes and appropriateness of intermediate care were presented. The chapter has also highlighted the statistical problems that were inherent in the demonstration dataset which are the subject of the empirical analyses in chapters five, seven and nine. The key messages for health services researchers working in this or other related fields were also outlined. The next chapter begins addressing the first of these problems, namely the challenges that are associated with the distributional characteristics of a dependent variable in multivariate analysis. Data from this national evaluation of costs and outcomes for intermediate care services for older people in the UK are used to illustrate the methods for dealing with these problems.

Table 4.1: Summary characteristics of the case-study sites

Site	Population range	Nature of area	Description of health and social care system	Description of intermediate care system	Types of intermediate care services provided
A	250,000 – 270,000	Urban	City-wide PCT, coterminous with social services	No single point of access for intermediate care. Services operating as a managed network	Rapid response/community teams Supported discharge teams (with domiciliary care) Residential intermediate care: nurse-led unit in acute hospital; community rehabilitation beds; sheltered housing facility
B	230,000 – 250,000	Urban	One of 4 PCTs covering large city. Social services city-wide	Single point of access for intermediate care with professionally qualified referral taker	Rapid response/community team Residential intermediate care: intermediate care unit and beds in independent sector care home)
C	150,000 – 170,000	Urban/semi-rural	One of 3 PCTs covering county. Working alongside county-wide social services	Single point of access for intermediate care with non-qualified referral taker	Rapid response rehabilitation teams Residential intermediate care: purpose built intermediate care unit; sheltered housing facility

Adopted from ICNET (2005)

Table 4.1: Summary characteristics of the case-study sites

Site	Population range	Nature of area	Description of health and social care system	Description of intermediate care system	Types of intermediate care services provided
D	170,000 - 190,000	Rural/semi-rural	One of 3 PCTs covering county. Working alongside county-wide social services	Two points of access for intermediate care (geographically based) with non-qualified referral takers	Rapid response Rehabilitation domiciliary care Rehabilitation units (day centre, day hospital, community hospital) Residential intermediate care: beds in independent sector residential home; sheltered housing facility
E	210,000 – 230,000	Urban and rural	One of 3 PCTs covering county. Working alongside county-wide social services	Single point of access for intermediate care with non-qualified referral taker	Community rehabilitation teams Day hospital (rapid assessment and rehabilitation) Residential intermediate care: beds in community facility; beds in independent residential home; sheltered housing facility

Adopted from ICNET (2005)

Table 4.2: Variables for use in economic analysis (with level of completeness)

Variable	Description	Missing (%)
Episode Characteristics		
Age	Age on 01/01/03	3
Gender	1 = female , 0 = Male	2
Live alone	1 = Individual lives alone, 0 = Otherwise	9
Barthel – Start	Barthel Score at start of IC episode	31
Barthel – End	Barthel Score at end of IC episode	38
EQ5D – Start	EQ-5D at start of IC episode	40
EQ5D – End	EQ-5D at end of IC episode	41
Change in ED-5D	Difference between EQ-5D score at end and at start of IC episode	42
Change in Barthel	Difference between Barthel score at end and at start of IC episode	41
Cost	Cost per patient	31
Descriptors of IC Services		
	<i>Type of service required</i>	3
Admission Avoidance service	1 = Acute Admission Avoidance service, 0 = Otherwise	
Supported Discharge service	1 = Supported discharge service, 0 = Otherwise	
Other Service	1 = Other IC Services, 0 = Otherwise	
Type of IC	1 = Residential IC, 0 = Non-Residential IC	0
	<i>Outcome of IC episode</i>	13
Transfer	1 = Transferred before end of IC episode, 0 = Other outcome	
Complete	1 = Completed IC episode, 0 = Otherwise	
Died	1 = Patient Died, 0 = Otherwise	
Other Outcome	1 = Alternative Outcome, 0 = Other outcome	
Stay Duration	Duration of service provision (number of days)	17
Descriptors of IC related services		
	<i>Source of referral</i>	3
Referral – primary	0 = Otherwise, 1 = Primary Care	
Referral – hospital	0 = Otherwise, 1 = Hospital	
Referral – social	0 = Otherwise, 1 = Social Services	
Referral – other	0 = Otherwise, 1 = Other Sources	
	<i>Alternatives to IC services</i>	18
Alternative – Home	0 = Else, 1 = Home	
Alternative – Hospital	0 = Else, 1 = Hospital	
Alternative – other	0 = Else 1 = Other alternative	

IC = Intermediate Care

Table 4.3: Baseline Patient Characteristics

Baseline characteristic	All patients (n = 2253)	Admission Avoidance (AA) patients (n = 1200)	Supported Discharge (SPD) patients (n = 909)	Statistical Tests of Difference between AA and SPD Patients
Age [Mean; Median (IQR)]	79; 81 (75, 87)	80; 82 (76, 87)	78; 80 (73, 85)	$\chi^2 = 39.9$ (p < 0.01) \pm
Gender [n Male (valid %)]	681 (30)	350 (30)	287 (32)	$\chi^2 = 1.5$ (p = 0.046) π
Lives alone [n (valid %)]	1096 (54)	579 (53)	477 (55)	$\chi^2 = 1.0$ (p = 0.6) π
In receipt of support services [n (valid %)]	512 (61)	302 (72)	193 (48)	$\chi^2 = 82.1$ (p < 0.01) \pm
EQ-5D on admission [Mean; Median (IQR)]	0.42; 0.52 (0.19, 0.69)	0.32; 0.43 (0.00, 0.62)	0.50; 0.59 (0.26, 0.71)	$\chi^2 = 67.3$ (p < 0.01) \pm
Barthel on admission – out of 20 [Mean; Median (IQR)]	14.8; 16.0 (12, 18)	13.8; 15.0 (11, 17)	15.6; 17.0 (14, 18)	$\chi^2 = 47.1$ (p < 0.01) \pm

\pm - Kruskal Willis test; π - Chi-square test

Table 4.4: Intermediate care episode costs and health changes (mean, SD)

	Total sample			
	All patients (n = 2253)	Admission Avoidance (AA) patients (n = 1200)	Supported Discharge (SPD) patients (n = 909)	Statistical Tests (\pm)- of Difference between AA and SPD Patients
Duration of Service provision in days (SD)	29.5 (31.2)	26.4 (26.8)	33.8 (34.0)	$\chi^2 = 33.3$ (p < 0.01)
Cost per patient in £ (SD)	1216 (1413)	980 (1165)	1581 (1712)	$\chi^2 = 110.6$ (p < 0.01)
EQ-5D on admission (SD)	0.42 (0.36)	0.32 (0.39)	0.50 (0.32)	$\chi^2 = 67.3$ (p < 0.01)
Barthel on admission (SD)	14.81(4.22)	13.85 (4.56)	15.63(3.72)	$\chi^2 = 47.1$ (p < 0.01)
Change in EQ-5D (SD)	0.16 (0.32)	0.23 (0.38)	0.12 (0.27)	$\chi^2 = 12.5$ (p < 0.01)
Change in Barthel (SD)	1.68 (2.89)	2.11 (3.31)	1.42 (2.59)	$\chi^2 = 3.7$ (p = 0.16)

\pm - Kruskal Willis test

Table 4.5: Intermediate care episode costs and health changes (mean, SD)

	All Patients			Admission Avoidance Patients			Supported Discharge Patients		
	Appropriate intermediate care (n = 974)	Inappropriate intermediate care (n = 1279)	Statistical test of difference (Kruskall Wallis test)	Appropriate intermediate care (n = 435)	Inappropriate intermediate care (n = 765)	Statistical test of difference (Kruskall Wallis test)	Appropriate intermediate care (n = 525)	Inappropriate intermediate care (n = 384)	Statistical test of difference (Kruskall Wallis test)
Duration of Service provision in days (SD)	27.4 (30.0)	31.2 (32.0)	$\chi^2 = 6.5$ (p = 0.01)	22.7 (24.6)	28.5 (27.8)	$\chi^2 = 13.6$ (p < 0.01)	30.8 (33.1)	38.5 (34.8)	$\chi^2 = 15.2$ (p < 0.01)
Cost per patient in £ (SD)	1457 (1711)	1018 (1141)	$\chi^2 = 39.3$ (p < 0.01)	1008 (1180)	966 (1157)	$\chi^2 = 1.3$ (p = 0.25)	1818 (1978)	1214 (1096)	$\chi^2 = 11.7$ (p < 0.01)
EQ-5D on admission (SD)	0.44 (0.34)	0.4 (0.38)	$\chi^2 = 1.6$ (p = 0.21)	0.27 (0.39)	0.34 (0.39)	$\chi^2 = 3.7$ (p = 0.053)	0.50 (0.31)	0.5 (0.36)	$\chi^2 = 0.2$ (p = 0.61)
Barthel on admission (SD)	14.47 (4.03)	15 (4.4)	$\chi^2 = 12.6$ (p < 0.01)	12.68 (4.30)	14.3 (4.6)	$\chi^2 = 19.0$ (p < 0.01)	15.15 (3.72)	16.5 (3.6)	$\chi^2 = 27.9$ (p < 0.01)
Change in EQ-5D (SD)	0.17 (0.31)	0.16 (0.32)	$\chi^2 = 0.69$ (p = 0.41)	0.28 (0.44)	0.21 (0.35)	$\chi^2 = 1.5$ (p = 0.22)	0.13 (0.27)	0.11 (0.28)	$\chi^2 = 3.1$ (p = 0.08)
Change in Barthel (SD)	1.97 (2.99)	1.4 (2.8)	$\chi^2 = 18.9$ (p < 0.01)	2.94 (3.72)	1.8 (3.1)	$\chi^2 = 8.3$ (p < 0.01)	1.69 (2.68)	0.9 (2.3)	$\chi^2 = 20.4$ (p < 0.01)

CHAPTER FIVE - CASE STUDY: A NATIONAL EVALUATION OF THE COSTS AND OUTCOMES OF INTERMEDIATE CARE SERVICES FOR OLDER PEOPLE – STATISTICAL MODELLING

5.1 Introduction

This chapter presents an opportunity to demonstrate how problems that are associated with the distributional characteristics of the dependent variable(s) in a regression model can be dealt with. The data used is from the national evaluation of costs and outcomes of intermediate care services for older people in the UK, which was described in chapter four. The objective of the regression analyses in this chapter was to determine what factors explain variation in both health and functional outcomes of older people between admission and discharge. The regression models therefore sought to understand which variables were significant predictors of the change in EQ-5D (health outcome) and change in Barthel index (functional outcome). The EQ-5D and Barthel index were described in detail and their distributional characteristics examined in the previous chapter. In this chapter, methods for dealing with the problems arising from these distributional characteristics are explored in a regression framework. The regression modelling approach chosen for this work is justified and a discussion of the results of the regression analysis follows. Principal messages for health services researchers dealing with similar datasets are also presented.

5.2 Method of analysis

A regression framework was used in the analysis of factors that may explain the change in measures of outcome. Regression modelling is an approach that uses a set of individual data points to describe the mathematical relationship between two or more variables (Gujarati 1995). In particular, the approach determines whether a

change in one variable causes a change in another (Wooldridge 2002). It was therefore a fitting approach to use to ascertain which variables caused changes in the outcomes of older people. Regression modelling explored the variability in health outcomes and was a means of introducing statistical control into the comparisons that were being made. The independent variables used in all models included patient characteristics, descriptors of the intermediate care service, features of the intermediate care episode and sources of referral (Table 5.1). This analysis did not address issues related to truncation, censoring or zero values of cost, which can be potential problems in any analysis (Jones 2000; Manning 1998; Mullahy, 1998; Blough et al. 1999). This was because these problems were not encountered in the analysis of the demonstration dataset. STATA version 8.2 (StataCorp LP, 2004) was used for all regressions.

5.3 Outcome Measures

The type of economic analysis to be conducted determines the kind of health or functional outcome measures to be used. The two measures chosen for this work were the EQ-5D and the Barthel Index. These outcomes were described in detail in chapter four. Data on the health and functional outcome variables were collected at two points: at admission to the intermediate care service and at discharge. The objective here was to find out what variables are predictors of the change in outcome measures between admission and discharge. The difference in the outcome variables between the two points was calculated. The dependent variables were therefore defined as follows:

$$(i) \text{ Change in EQ-5D } (\Delta EQ\text{-}5D) = EQ5D2 - EQ5D1$$

$$(ii) \text{ Change in Barthel } (\Delta Barthel) = Barthel2 - Barthel1$$

where:

EQ5D1 = EQ5D score at admission; EQ-5D2 = EQ5D score at

discharge;

Barthel1 = Barthel score at admission and Barthel2 = Barthel score at

discharge.

5.4 Common solutions for dealing with problems associated with distributional characteristics of variables

5.4.1 Ceiling Effects

Due to the ceiling effect in the EQ-5D and Barthel, some studies have suggested that ordinary least squares (OLS) methods may not perform as well as other studies that take into account this ceiling effect particularly when predictive rather than explanatory models are considered (Brennan and Spencer, 2006; Sullivan & Ghushchyan 2006a). Consequently, some studies that have analysed quality of life data of this kind have used truncated regressions and censored least absolute deviations (CLAD) regressions (Sullivan & Ghushchyan 2006a, Clarke et al. 2002, Saarni et al. 2006; Payakachat et al. 2009). Tobit models have also been suggested as a means of accounting for these effects (Greene 1997 and Long 1997).

5.4.2 Skewness

A common method for dealing with skewness is log-transforming the dependent variable i.e. taking the natural log of the variable (Manning & Mullahy 2001; Manning 1998; Mullahy 1998). But this method has several disadvantages: (i) the

difficulty associated with interpreting the coefficients because they are on a different scale to the original, (ii) the retransformed parameter estimates are usually of poor quality unless they are appropriately retransformed (e.g. using Duan 1983 retransformation method) and (iii) the transformed data will usually only have an approximate normal distribution (Cantoni & Ronchetti 2004; Duan 1983). A method that addresses the problem of skewness without the three disadvantages discussed above is the generalised linear model (GLM) framework (McCullagh & Nelder 1989). This method has the added advantage of jointly dealing with the problem of robustness as well as that of heteroscedasticity. Barber and Thompson (2000) have also suggested bootstrapping in the context of a randomised trial.

5.4.3 Heteroscedasticity

A common solution for dealing with this problem is the use of standard errors in the regression equations estimated from White's heteroscedasticity consistent covariance matrix (White 1980). Heteroscedastic transformation has also been suggested (Zhou et al. 2008). Manning and Mullahy (2001) proposed another method for dealing with the twin problem of heteroscedasticity and skewness. It involves running an OLS model for the natural log of dependent variable and then using a non-graphical test to check for heteroscedasticity. But since there is the presence of non-positive values in our outcome measures, a shift technique (i.e. adding a constant until all values are non-zero) had to be used. If heteroscedasticity is detected from the OLS residuals on a log-scale, then heteroscedastic retransformation has to be done or one of the GLM models used.

5.5 Regression Models

The regression models chosen reflect the need to deal with the problems associated with the distributional characteristics of the dependent variables.

5.5.1 Regression model specification

The regression specification depicts the relationship between the dependent variables, which are the quality of life and functional independence outcomes, and the explanatory variables. The quality of life outcome is expressed as the change in EQ-5D ($\Delta\text{EQ-5D}$) while functional independence is expressed as the change in the Barthel index ($\Delta\text{Barthel}$). Algebraically, this relationship can be depicted as:

$$O_i = f(X_i) + \mu_i \quad (5.1)^4$$

where O_i denotes the measure of outcome ($\Delta\text{EQ-5D}$ or $\Delta\text{Barthel}$) for the i th individual,

X_i is a vector of explanatory variables for the i th individual and

μ_i is the stochastic error term for the i th individual.

Three regression model families were considered and all were run. These were the CLAD model, the GLM and OLS. The results from the OLS and CLAD models are discussed in sections 5.5.1 and 5.6, respectively.

5.5.2 CLAD model Specification

The CLAD model depicted as follows:

⁴ $f(X_i)$ means 'some function of the variable X_i '

$$O_i^* = f(X_i) + \mu_i \quad (5.2)$$

where:

$O_i^* = \Delta EQ-5D_i^*$ or $\Delta Barthel_i^*$ which are the individual i 's true change in EQ-5D score or change in Barthel score, respectively, and
 X_i = vector of individual i 's characteristics.

Then $O_i = \Delta EQ-5D_i = 1.59$ if $\Delta EQ-5D_i^* \geq 1.59$ and

$$\Delta EQ-5D_i = \Delta EQ-5D_i^* \text{ if } \Delta EQ-5D_i^* < 1.59.$$

$O_i = \Delta Barthel_i = 20$ if $\Delta Barthel_i^* \geq 20$ and

$$\Delta Barthel_i = \Delta Barthel_i^* \text{ if } \Delta Barthel_i^* < 20$$

The CLAD model is similar in specification to the Tobit model (Greene 1997). However, whereas the Tobit model would characterise the relationship between the mean change in EQ-5D (Barthel) and X , the CLAD model describes the association between the median change in EQ-5D (Barthel) and X (Huang et al. 2008).

5.5.3 GLM model specification

The GLM can be broken down into 3 parts: the linear component, a differentiable link function and a variance function (Blough et al. 1999). The relationship between outcomes and the explanatory variables can then be expressed as:

$$O_i = f(X_i) + \mu_i \quad (5.3)$$

$$E(O/X_i) = g^{-1} f(X_i) \quad (5.4)$$

where:

$E(O/X_i)$ = is the conditional mean of $\Delta EQ-5D_i$ or $\Delta Barthel_i$ given X_i and

g = monotonic differentiable link function describing how $E(O/X_i)$ is related to $f(X_i)$

5.5.3.1 Choosing an Estimator

To select the particular GLM model to use, Manning and Mullahy (2001) suggested an algorithm which employs the raw scale residuals from a GLM with a log link. A squared raw scale residual in a modified park test is used to determine the appropriate family (distribution) to employ from among the GLM alternatives (ibid; Manning et al. 2001).

Characterizing the class of variance functions by:

$$Var(O/X_i) = \delta_i^2 = \alpha [E(O/X_i)]^\lambda \quad (5.6)$$

where:

$Var(O/X_i)$ or δ_i^2 = variance of $\Delta EQ-5D_i$ or $\Delta Barthel_i$ given X_i ,

α = dispersion factor,

λ = finite and non negative factor.

Provided that λ is finite and non negative, then the choice of GLM distribution is guided by:

$\lambda = 0 \Rightarrow$ usual non-linear least squares estimator.

$\lambda = 1 \Rightarrow$ Poisson like class.

$\lambda = 2 \Rightarrow$ Gamma, Homoscedastic Log-normal, Weibull distribution.

$\lambda = 3 \Rightarrow$ Inverse Gaussian (Wald) distribution.

Blough et al (1999) proposed the use of a measure of goodness-of-fit in terms of the scaled deviance to decide on the best GLM model to use.

Though the GLM model provided estimates in original metric which did not require any retransformation before interpretation (Manning & Mullahy 2001; Kilian et al. 2002), exponentiated coefficients were used because of ease in interpretation (Kilian et al. 2002). In general, the deviance and the log-likelihood statistic are used to estimate the overall fit of a generalised linear model. In the models whose results are reported in this chapter, the more familiar R^2 was used. The R^2 was computed as the square of the correlation between the outcomes (the dependent variables) and the predicted values of the outcomes as suggested by Zheng and Agresti (2000).

5.5.4 OLS Model Specification

The OLS model can be depicted as follows:

$$O_i = f(X_i) + \mu_i \quad (5.7)$$

$$E(O/X_i) = f(X_i) \quad (5.4)$$

where:

$E(O/X_i)$ = is the conditional mean of $\Delta EQ-5D_i$ or $\Delta Barthel_i$ given X_i and

μ_i = is as defined in equation 5.1

5.5.5 Regression Diagnostics

To prevent erroneous or biased results, the models were checked for specification error, goodness-of-fit, multicollinearity and outliers (Gujarati, 1995).

5.5.6 Hypothesised relationships

No a-priori assumptions about the relationships between the variables were made. This underlined the exploratory nature of these regression analyses. As a result, Stepwise regression was employed to determine which variables were to be included in the regression models (Draper & Smith 1998).

5.5.7 Missing data

Complete case analysis was conducted in this chapter. Therefore, only variables with complete observations were used in the regression models. As a result, 584 observations were used in the model where $\Delta EQ-5D$ was the dependent variable while 607 observations were utilised in the model with $\Delta Barthel$ as the dependent variable. Methods of dealing with missing data are the subject of chapters six and seven.

5.6 Results

The regression models with the natural logs of the EQ-5D and Barthel differences as dependent variables were run and then the White test performed. Results showed the presence of heteroscedasticity ($\chi^2 = 11$, $p < 0.01$ and $\chi^2 = 37$, $p < 0.001$ for the EQ-5D

and Barthel, respectively). The algorithm suggested by Manning and Mullahy (2001) was then used to decide which GLM model to use. The GLM models for EQ-5D and Barthel yielded coefficients of 1.717 and 1.727, which are both closest to 2. These results suggested the Gamma distribution as the right choice of distributional family of the GLM model for both outcomes. A log link function was used for both models.

5.6.1 Interpreting GLM regression results

The results of the regression analyses are shown in Tables 5.2 and 5.3. As indicated in Table 5.1, there was a mixture of continuous and categorical variables among the dependent and independent variables. The value of a coefficient of a continuous independent variable in a regression model shows the amount by which the dependent variable will change due to a change of one unit in the independent variable. The value of a coefficient of a categorical independent variable in a regression model shows the amount by which the dependent variable will change due to a change in the independent variable from one category, level or class to another.

The values of exponentiated coefficients [$\text{Exp}(\text{coeff})$] show the proportionate magnitude by which $\Delta\text{EQ-5D}$ or $\Delta\text{Barthel}$ change for a unit change in the independent variables (Kilian et al. 2002). Table 5.2 shows that $\Delta\text{EQ-5D}$ is influenced by the EQ-5D score at admission, the Barthel score at admission and whether or not an intermediate care service admission was appropriate. The values of the exponentiated coefficients show that an increase in the EQ-5D score at admission of one unit causes a decrease in $\Delta\text{EQ-5D}$ of approximately 40% while an increase in the Barthel score at admission of one unit causes an increase in $\Delta\text{EQ-5D}$ of approximately 1%. Patients for whom an intermediate care admission was appropriate had a 6% bigger $\Delta\text{EQ-5D}$

than those for whom an admission was not appropriate. The value of 0.304 for R^2 shows that only 30.4% of the variation in $\Delta EQ-5D$ could be explained by the model. The results from Table 5.3 show that $\Delta Barthel$ is influenced by age, Barthel score at admission, whether an individual was transferred before the end of an intermediate care episode or if he/she had another intermediate care outcome, and whether an individual was referred by a primary care source or by social workers. From the values of the exponentiated coefficients, an increase in age of one year was associated with a 3% drop in $\Delta Barthel$ while an increase of one unit in the Barthel score at admission led to a decrease in $\Delta Barthel$ of 34%. The $\Delta Barthel$ for an individual who was transferred before the end of an intermediate care episode was about 99% lower than that of an individual who had another intermediate care outcome and an individual referred by a primary care source had a $\Delta Barthel$ which was about 80% lower than that of an individual 'referred' by social workers. The value of R^2 was 0.308 which shows that only 30.8% of the variation in $\Delta Barthel$ could be explained by the model.

5.6.2 Comparison to OLS results

To ascertain the necessity of undertaking the GLM approach, the results in 5.5.1 were compared against those obtained using an OLS model that did not account for skewness or heteroscedasticity. Of interest was whether or not the two models would lead to different results being obtained when used on the same variables. The key objective of the regression models was to find out which factors were significant predictors of change in the outcome variables. The results from the OLS model are depicted in tables 5.4 and 5.5. The interpretation of the OLS coefficient values is

broadly the same as that explained in section 5.6.1 above. Here, however, the values of the coefficients were not exponentiated as is usual practice (Gujarati, 1995).

In the Δ EQ-5D OLS model, five variables were found to be significantly related to change in EQ-5D as opposed to only three in the Δ EQ-5D GLM model. In the OLS model therefore, more variables significantly affected the outcome variable. Of these five, only two (EQ5D at admission and Appropriate intermediate care – IC) were significant in both models. The ‘Barthel at admission’ variable was only significant in the Δ EQ-5D GLM model while ‘Lives alone’, ‘Transferred before end of intermediate care’ and ‘Referral – Other’ were only significant in the Δ EQ-5D OLS model. This would lead to different conclusions being reached about which variables were significant in explaining variation in the Δ EQ-5D variable.

In the Δ Barthel OLS model, six variables, two more than in the Δ Barthel GLM model, were found to be significantly related to Δ Barthel. Of the six, five (Lives alone, EQ5D at admission, Appropriate intermediate care – AIC, Type of intermediate care and Completed intermediate care episode) were previously not significant. Only one (Barthel at admission) was significant in both models. Here too, it is evident that different results and thereafter different conclusions about which variables were significant in explaining variation in the Δ Barthel variable would have been reached if the OLS model was used instead of the GLM approach.

The key objective of running the regression models for the two outcome measures was to identify the particular variables that would significantly explain variation in Δ EQ-5D and Δ Barthel. Determining the specific variables that significantly cause

changes in the outcome measures is therefore important and central to this analysis. Identification of these particular significant variables would guide health services researchers when trying to understand what variables drive movement in the Δ EQ-5D and Δ Barthel variables. The fact that, for both sets of models, more variables were found to be significant in the OLS models and also that some previously significant variables in the GLM model were no longer significant in the OLS models therefore implies that different results would be arrived at when using the OLS model instead of the GLM approach. And as already established in section 4.10.4.3, both outcome measures were found to have skewed distributions while section 4.10.4.4 showed that the OLS models for the two outcome measures were heteroscedastic. Using the OLS therefore in the face of these statistical problems would cast a lot of doubt on the validity of the results obtained as the estimators are likely to be inefficient and the means, standard deviations and standard errors biased (Gujarati, 1995). Therefore, compared to the OLS, the GLM model can be viewed as better on theoretical grounds.

5.6.3 Key results from the analyses

Combining the messages obtained from the analysis carried out in chapter four and GLM regression analysis in this chapter will help give a fuller picture of the relationship between the outcomes (Δ EQ-5D and Δ Barthel) and other variables in the dataset.

The very positive message from the results reported in chapter four was that intermediate care services, on average, were associated with gains in health, indicated by a positive mean improvement in both the EQ-5D and the Barthel scores from admission to discharge. This was seen for the whole data set but appeared to be larger

for patients admitted for admission avoidance rather than supported discharge. However, it was unclear whether this larger gain for admission avoidance patients was intrinsic to the intermediate care service or resulted from the fact that admission avoidance service tended to be caring for more ill patients, with greater potential for health improvement.

The GLM regression analysis results, reported in Tables 5.2 and 5.3, shed some light on this as they provided estimates of health and functional gains by intermediate care function, with statistical control for potential confounding factors, such as patient severity. The regression model indicated that the difference in health and functional gains seen between admission avoidance and supported discharge services in chapter four was not statistically significant. Thus, the health and functional gains of intermediate care patients in the demonstration dataset appeared to be driven by illness severity at admission (the more severely ill patients gained most) rather than by intermediate care service function.

These results suggested that there were significant and important gains in functioning (as measured by Barthel) and in health-related quality of life (as measured by EQ-5D) delivered by the intermediate care services. The results also had some implications for policy in relation to which patients and what type of services were supposed to be given priority. The data suggested that there was little to be gained from the selection of patients on the basis of some socio-demographic characteristics (like gender). However, the need or severity of the patients appeared to be an important factor – those most likely to benefit from intermediate care were those with the greatest functional deficiencies measured using the Barthel functional measure and those with

lowest quality of life, measured by the EQ-5D. This focus on need was supported by the finding that greater benefits appear to be associated with patients for whom the intermediate care admission was deemed appropriate, that is, where the alternative to intermediate care would have involved a hospital episode.

The work reported here had a number of limitations. Certain services had missing data on the Barthel and EQ-5D. This was due to either logistical difficulties or the fact that it was not their policy to routinely collect the data. This meant a reduction in the potential sample size for modelling outcomes. The issue of missing data is the focus of chapters six and seven. In addition, the low R-squared values in the regression models highlighted limitations in the explanatory power of the models. Although a model with a low R-squared is still helpful in explaining variation in the outcome variable (Zheng & Agresti 2000, Schemper 2003), there was scope for improving the model fit of the regression models by, for instance, using a much wider selection of independent variables. Further, in explaining variation in outcomes, the intention was to differentiate between service configurations by dividing the services into the type of setting (residential or non-residential) or the type of service provided (admission avoidance, supported discharge or other). Therefore, this analysis did not differentiate between case study sites. The ceiling and floor effects were potentially problematic if they had not been taken into account. Powell's censored least absolute deviations (CLAD) estimator regression taking the ceiling effect into account was used as a way of testing whether the results of the GLM were valid. Using CLAD did not affect the results appreciably. In choosing the GLM over the CLAD model, more weight was given to the need to tackle the twin problems of skewness and heteroscedasticity (using the GLM) as opposed to solving the twin problems of ceiling/floor effects and

heteroscedasticity (using the CLAD model). This was because this trade-off was deemed alright as the focus of the regression models was on explaining variation in the outcome measures and not obtaining their predictions. If the focus was the latter, it would have been more important to ensure that the outcome measures were constrained between the permitted lower and upper limits. This is the case in chapter nine where the focus of the core analysis is on prediction. Tobit models could have been employed here but distributional assumptions for this approach were violated as has also been shown elsewhere (Austin 2002).

5.7 Discussion

5.7.1 Key findings

Why the GLM model was worth running

An examination of the Δ EQ-5D and Δ Barthel variables in chapter four showed that they were skewed and heteroscedastic. Using robust methods for dealing with these variables in a regression framework was therefore important. Various methods from the literature for undertaking this kind of analysis under these circumstances were identified and presented in section 5.4. In this chapter, the GLM was used to model change in EQ-5D and change in Barthel scores. This regression model was chosen because it is robust to both skewness and heteroscedasticity. However, to ascertain what effect a GLM approach would have on the regression analysis, the results of a simple OLS that does not take into account the problems of heteroscedasticity and skewness were also compared.

What has been shown?

As demonstrated in section 5.6.2, the GLM model produced results that were fairly different to those of the OLS (which did not account for skewness and heteroscedasticity) concerning which variables were significant predictors of change in the outcome variables. In the OLS models, more variables were found to be significantly related to the outcome variables while some variables which were significant in the GLM model were found not to be significant in the OLS models. Further, the diagnostic tests for the GLM also show that the resultant model was a good fit (i.e. the predicted deviance residuals were normally distributed - McCullagh and Nelder, 1989) and this coupled with the theoretical advantages of the GLM in the face of skewness and heteroscedasticity gave more validity to the GLM results as opposed to those from the OLS.

Why does it matter which model is used?

Outcome data such as EQ-5D and Barthel are most likely to be always skewed with the resultant residuals being heteroscedastic. Robust methods of dealing with these problems related to the distributional characteristics of the variables will always be invaluable in order to obtain unbiased results. When the main objective of a regression analysis is to identify which specific variables are important in explaining variation in an outcome measure, the choice of the regression method to deal with problems associated with the distributional characteristics of the measure is not without consequence. This is because different methods can result in completely different independent variables being shown to be significantly related to the outcome measure thereby leading to different conclusions concerning which factors are key predictors of the measure. The fact that fairly different results and conclusions could

have been reached about which variables were significant in explaining variation in the outcome variables if the GLM was not used shows that the method used to deal with skewed and heteroscedastic dependent variables matters.

5.7.2 Key messages for health services researchers from this chapter

The goal of the regression analysis reported in this chapter was to find out which variables significantly explain variation in the two outcome measures (Δ EQ-5D and Δ Barthel). These two outcome measures were shown in chapter four to be skewed, heteroscedastic and also to have had a ceiling effect. As a result, three models (GLM, CLAD and OLS) were run and their results compared. Because the goal of the regression analyses was to explain variation in the outcome measures as opposed to predicting them, a decision was made to focus on the twin problems of heteroscedasticity and skewness and not on the ceiling effect (a statistical problem for which the CLAD model has been shown to be robust to, as will be discussed further in chapter nine). Therefore emphasis was not placed on comparing the CLAD to the GLM (even though the results from the two models were broadly the same) but on a comparison between the GLM and the OLS models.

Determining the specific variables that are significantly responsible for changes in the two outcome measures was central to this analysis as it would guide health services researchers when trying to understand what variables drive movement in the Δ EQ-5D and Δ Barthel variables. Since more variables were found to be significant in the OLS models and also that some previously significant variables in the GLM model were no longer significant in the OLS models, different results and therefore conclusions could be arrived at if the OLS model was used instead of the GLM approach. A choice must

therefore be made between the two models. In chapter four, the two outcome measures were shown to be skewed and heteroscedastic. The GLM approach has been previously shown to be theoretically robust to these two problems (Gujarati, 1995; Manning & Mullahy, 2001) and therefore a solution from this model can be viewed as better theoretically to that from the OLS.

5.8 Conclusion

This chapter has shown how to deal with the problems that arise because of the distributional characteristics of variables using a GLM model in a regression framework. What the analyses carried out also highlighted is that more than one method is always at the disposal of an analyst faced with such a problem. However, careful consideration must be taken to ensure that the method chosen does not violate the assumptions behind it. The use of a model that does not account for skewness or heteroscedasticity such as a simple OLS will lead to erroneous results being obtained.

The next chapter moves to the second of our statistical issues: missing data. Again data from the national evaluation of costs and outcomes of intermediate care for older people in the UK were used to demonstrate the effects of making different assumptions about the missingness when choosing the statistical approach to use for coping with the missing data.

Table 5.1: Variables for use in economic analysis

Variable	Description
Episode Characteristics	
Age	Age in years on 01/01/03 (continuous variable)
Gender	Dummy variable coded 1 if individual is female and 0 if male
Live alone	Dummy variable coded 1 if Individual lives alone and 0 otherwise
Barthel – admission	Barthel Score at admission (continuous variable)
Barthel – discharge	Barthel Score at discharge (continuous variable)
EQ5D – admission	EQ-5D at admission (continuous variable)
EQ5D – discharge	EQ-5D at discharge (continuous variable)
Change in ED-5D	Difference between EQ-5D score at discharge and at admission (continuous variable)
Change in Barthel	Difference between Barthel score at discharge and at admission (continuous variable)
Descriptors of intermediate care (IC) Services	
Admission Avoidance Service	<i>Type of service required</i> Dummy variable coded 1 if service was Acute Admission Avoidance and 0 if not
Appropriate IC	Dummy variable coded 1 if IC service admission was appropriate and 0 if it was not
Type of IC	Dummy variable coded 1 if IC was in a residential setting and 0 if it was not
Transfer	<i>Outcome of IC episode</i> Dummy variable coded 1 if an individual was transferred before end of IC episode and 0 otherwise
Complete	Dummy variable coded 1 if an individual completed an IC episode and 0 otherwise
Died	Dummy variable coded 1 if a patient died and 0 if he/she did not
Other Outcome	Dummy variable coded 1 if an individual had an alternative outcome and 0 otherwise
Stay Duration	Duration of service provision - number of days (continuous variable)
Descriptors of IC related services	
Referral – primary	<i>Source of referral</i> Dummy variable coded 1 if an individual was referred by a primary care source and 0 otherwise
Referral – hospital	Dummy variable coded 1 if an individual was referred by a hospital and 0 otherwise
Referral – social	Dummy variable coded 1 if an individual was ‘referred’ by social workers 0 otherwise
Referral – other	Dummy variable coded 1 if an individual was referred by another service and 0 otherwise
Alternative – Home	<i>Alternatives to IC services</i> Dummy variable coded 1 if the alternative was a home and 0 otherwise
Alternative – Hospital	Dummy variable coded 1 if the alternative was a hospital and 0 otherwise
Alternative – other	Dummy variable coded 1 if the alternative was another place and 0 otherwise

Table 5.2: Change in EQ-5D Model – GLM model

Variables	Coeff.	Std err	Z	Exp (Coeff)
Age	0.000	0.001	0.010	1.000
Gender (1=female, 0=male)	0.032	0.025	1.290	1.032
Lives alone (1=lives alone, 0=Otherwise)	0.029	0.022	1.270	1.029
EQ5D at admission	-0.515	0.036	-14.370	0.598**
Barthel at admission	0.011	0.003	3.230	1.011**
Admission Avoidance Service -AAS (1= AAS, 0=Otherwise)	0.033	0.040	0.830	1.034
Appropriate intermediate care – AIC (1 = AIC, 0 = Otherwise)	0.058	0.025	2.310	1.060**
Type of intermediate care (1=residential, 0=Otherwise)	0.033	0.029	1.170	1.034
Duration of Service provision	0.000	0.000	-0.890	1.000
Transferred before end of intermediate care episode	-0.140	0.088	-1.590	0.869
Completed intermediate care episode	-0.003	0.064	-0.040	0.997
Other intermediate care Outcome (Reference. Group)				
Referral – Primary	0.003	0.053	0.060	1.003
Referral – Hospital	0.025	0.058	0.440	1.026
Referral – Other	0.139	0.085	1.640	1.149
Referral – Social Workers (Reference Group)				
Constant	0.111	0.131	0.850	1.117
R-Squared		0.304		

Dependent variable = EQdiff (Change in EQ-5D); n = 584

* Sig. at 5% level ** Sig. at 1% level

Table 5.3: Change in Barthel Model – GLM model

Variables	Coeff.	Std err	Z	Exp (Coeff)
Age	-0.030	0.012	-2.530	0.970**
Gender (1=female, 0=male)	-0.055	0.284	-0.190	0.947
Lives alone (1=lives alone, 0=Otherwise)	0.111	0.265	0.420	1.118
EQ5D at admission	0.450	0.469	0.960	1.568
Barthel at admission	-0.416	0.045	-9.200	0.660**
Admission Avoidance Service -AAS (1= AAS, 0=Otherwise)	1.067	0.599	1.780	2.906
Appropriate intermediate care – AIC (1 = AIC, 0 = Otherwise)	0.475	0.346	1.370	1.608
Type of intermediate care (1=residential, 0=Otherwise)	-0.357	0.354	-1.010	0.700
Duration of Service provision	-0.003	0.004	-0.940	0.997
Transferred before end of intermediate care episode	-4.494	1.664	-2.700	0.011**
Completed intermediate care episode	-0.698	1.575	-0.440	0.498
Other intermediate care Outcome (Reference Group)				
Referral – Primary	-1.615	0.787	-2.050	0.199*
Referral – Hospital	-0.128	0.830	-0.150	0.880
Referral – Other	-1.197	1.648	-0.730	0.302
Referral – Social Workers (Reference Group)				
Constant	6.371	1.550	4.110	584.390
R-Squared		0.308		

Dependent variable = Bartdiff (Change in Barthel Index); n = 607

* Sig. at 5% level ** Sig. at 1% level

Table 5.4: Change in EQ-5D Model – OLS Model

Variables	Coeff.	Std err	Z
Age	-0.001	0.000	-1.490
Gender (1=female, 0=male)	0.015	0.013	1.230
Lives alone (1=lives alone, 0=Otherwise)	0.026**	0.010	2.610
EQ5D at admission	-0.182**	0.014	-13.320
Barthel at admission	0.001	0.001	1.140
Admission Avoidance Service -AAS (1= AAS, 0=Otherwise)	0.027	0.016	1.720
Appropriate intermediate care – AIC (1 = AIC, 0 = Otherwise)	0.056**	0.013	4.200
Type of intermediate care (1=residential, 0=Otherwise)	0.017	0.012	1.420
Duration of Service provision	0.000	0.000	-0.660
Transferred before end of intermediate care episode	-0.136**	0.058	-2.340
Completed intermediate care episode	-0.001	0.026	-0.030
Other intermediate care Outcome (Reference. Group)			
Referral – Primary	0.013	0.016	0.810
Referral – Hospital	0.019	0.019	0.980
Referral – Other	0.102**	0.026	3.880
Referral – Social Workers (Reference Group)			
Constant	0.827	0.049	16.910
R-Squared		0.256	

Dependent variable = EQdiff (Change in EQ-5D); n = 584

* Sig. at 5% level ** Sig. at 1% level

Table 5.5: Change in Barthel Model – OLS model

Variables	Coeff.	Std err	Z
Age	-0.001	0.001	-1.140
Gender (1=female, 0=male)	-0.048	0.028	-1.750
Lives alone (1=lives alone, 0=Otherwise)	0.073**	0.026	2.830
EQ5D at admission	-0.112**	0.035	-3.200
Barthel at admission	-0.030**	0.003	-9.980
Admission Avoidance Service -AAS (1= AAS, 0=Otherwise)	0.029	0.038	0.760
Appropriate intermediate care – AIC (1 = AIC, 0 = Otherwise)	0.133**	0.030	4.470
Type of intermediate care (1=residential, 0=Otherwise)	0.060**	0.027	2.260
Duration of Service provision	0.000	0.000	-1.070
Transferred before end of intermediate care episode	-0.172	0.158	-1.090
Completed intermediate care episode	0.257*	0.127	2.020
Other intermediate care Outcome (Reference Group)			
Referral – Primary	0.069	0.049	1.390
Referral – Hospital	-0.006	0.054	-0.110
Referral – Other	0.118	0.077	1.530
Referral – Social Workers (Reference Group)			
Constant	1.332**	0.164	8.140
R-Squared		0.242	

Dependent variable = Bartdiff (Change in Barthel Index); n = 607

* Sig. at 5% level ** Sig. at 1% level

CHAPTER SIX – CHALLENGES WITH MISSING DATA: THEORY

6.1 Introduction

The previous chapter dealt with the first of the three prevalent statistical challenges identified in chapter three and also present in the demonstration dataset, namely problems arising from the distributional characteristics of dependent variables. In this chapter, the attention moves to the second of these problems, missing data, and the approaches that can be used to deal with it. Missing data is an unwanted reality in most evaluations as it can lead to threats to the internal and external validity of the results obtained from analysing such data (Schafer 1997; Kline 1998; Chen & Shao 2000). This chapter tackles the theory around this problem by focussing on the three mechanisms that may be responsible for missing data. The mechanisms are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The chapter also considers the methods that can be used to cope with missing data under each of the three mechanisms. Ways of determining the mechanism behind missing data are also addressed. The key messages for health services researchers are summarised before the conclusion to the chapter is given.

6.2 Adverse effects of missing data

There is a possibility that even under the best of conditions, missing data is likely to lead to a significant reduction in sample size. A sample reduced in size may no longer be representative of the target population thereby affecting its external validity as it will be difficult to generalise the results obtained from analysing such a sample to the a larger population (Kline 1998; Chen & Shao 2000; Rubin 1987). This is more of a problem in circumstances where the likelihood of response is related to observed

characteristics. Certain forms of missingness may reduce the statistical power of the analyses of the available data and therefore compromise the internal validity of a study, which is more serious (Little & Rubin 1987, 2002). A situation that can potentially lead to reduced internal validity is when the likelihood of response is related to the values of the variable for which values are only partially observed, which seems possible in a lot of cases.

There are various methods that have been proposed to deal with missing data (Allison 2000; Schafer 1999; Schafer 1997). For each of these methods, different assumptions about the missing data mechanism are made. Croninger and Douglas (2005) indicate that the method used for coping with missing data is not as important if there is not much data missing and/or the sample is big. This is because most strategies will yield similar results in such circumstances. But as the level of missingness rises and/or the sample becomes smaller, the choice of method becomes potentially more significant. However there are no clear rules concerning what is considered to be too little or too much missing data (Kline 1998). Cohen and Cohen (1983) considered 5 to 10% missing data on a variable to be small, while variables with 40% missing data were considered to be high (Raymond & Roberts 1987). Criteria for deciding the most appropriate method to be used under what circumstances need to be established. But even if appropriate strategies for coping with missing data are utilised, various methods may yield substantially different conclusions (Cohen et al. 2003).

6.3 Mechanisms that lead to missing data

Taxonomy of missing data mechanisms describing the process that generates the missing data has been explicated by many (Rubin 1976; Little & Rubin 1987, 2002;

Schafer 1997; Leon et al. 2006; Kline 1998). There are three main classes: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). If data are MAR or MCAR, they can also be referred to as “ignorable” data while those MNAR are “non-ignorable” (Foster & Fang 2004). Missing data are said to be ignorable if the parameters that are used to model the missing data process are not related to the parameters used to model the observed data while non-ignorability exists if there is a systematic difference between responders and nonresponders even after accounting for all the observed data (Kmetz et al. 2002; Croninger & Douglas 2005).

For notational purposes, let Y be a complete $n \times p$ matrix of complete data with elements y_{ij} . The i th row is then $y_i = (y_{i1}, \dots, y_{ip})$ where y_{ij} is the value of Y for subject i . The observable part of Y is represented by $Y_{observable}$ and the missing part by $Y_{missing}$. This means that Y can be defined as

$$Y = (Y_{observable}, Y_{missing}) \quad (6.1)$$

Following Rubin’s (1976) definition, we can further characterise $R = R_{ij}$ to be an $n \times p$ matrix of indicator variables which is defined as

$$R = R_{ij} = \begin{cases} = 0 & \text{if } Y = Y_{observable} \\ = 1 & \text{if } Y = Y_{missing} \end{cases} \quad (6.2)$$

Because R would be expected, in general, to be related to Y , we can define a probability model for R as follows:

$$P(R) = P(R | Y, V), \quad (6.3)$$

where V represents some unknown parameters

MCAR implies that the probability of an item missing is unrelated to any measured or unmeasured characteristic for that unit. The missing values for the variable y ($Y_{missing}$) are not related either to the variable itself or to any other variable in the matrix of complete data (Y). If this assumption holds, cases with valid values for Y ($Y_{observable}$) can be analysed with reasonable likelihood of being representative subsets of the original population. Data that are MCAR show the highest degree of randomness and do not present any underlying reasons for missing observations that can potentially bias research conclusions (Musil et al. 2002). An example of MCAR is if respondents accidentally miss out a question on a questionnaire or if data are missing due to some administrative reason, such as omissions at data entry. Suppose two variables, age and EQ-5D, are measured. Data can be said to be MCAR if the probability that EQ-5D is missing is the same for all individuals, regardless of age or EQ-5D i.e. the probability of having missing observations on the EQ-5D is independent of the EQ-5D score and also of age.

The MCAR assumption holds if:

$$P(R=1 | Y_{observable}, Y_{missing}, V) = P(R=1 | V) \quad (6.4)$$

or

$$f(R | Y_{observable}, Y_{missing}, V) = f(R | V), \text{ for all } Y \text{ and } V \quad (6.5)$$

MAR data exhibit some randomness to the pattern of data omission. If data are MAR, then missingness on Y ($Y_{missing}$) is related to another variable ($Y_{observable}$) in the analysis but not to $Y_{missing}$ itself. Therefore, the probability of an item having incomplete data depends on other variables in the dataset (Schafer 1997). This is the assumption underlying most imputation methods, since they use the observed data to predict what is missing (Schafer 1997). Using our age and EQ-5D example again, data can be said to be MAR if the likelihood of EQ-5D missing varies according to the age of a respondent but not according to the EQ-5D of a respondent of a given age i.e. the probability of having missing observations on the EQ-5D is independent of the EQ-5D score but is related to age.

The MAR assumption holds if:

$$P(R = I \mid Y_{observable}, Y_{missing}, V) = P(R = I \mid Y_{observable}, V) \quad (6.6)$$

or

$$f(R \mid Y_{observable}, Y_{missing}, V) = f(R \mid Y_{observable}, V), \quad (6.7)$$

for all $Y_{missing}$ and V

MNAR is when the probability of missingness depends on both the values of $Y_{missing}$ and $Y_{observable}$ as well as the values of one or more other variables (V) in the analytic model (Fielding et al. 2004). This implies that the missing observations would, if measured, have a different distribution for the predicted and for the observed. It is not possible to correct data for a MNAR mechanism, except by using outside information. Data can be said to be MNAR if the probability that the EQ-5D is missing varies

according to the EQ-5D and also according to age i.e. i.e. the probability of having missing observations on the EQ-5D is dependent on the EQ-5D score and also on age.

The MNAR assumption holds if:

$$P(R = I | Y_{observable}, Y_{missing}, V) = P(R | Y_{missing}, Y_{observable}, V) \quad (6.8)$$

or

$$f(R | Y_{observable}, Y_{missing}, V) = f(R | Y_{missing}, Y_{observable}, V), \quad (6.9)$$

for all Y and V

Leon et al (2006) discuss several other variations of missing data mechanisms. They present covariate-dependent MCAR (Little 1995) as a special case of MCAR in which missingness is associated with observed covariates and measured before dropout. Ignorable nonresponse refers to a special case of MAR where missingness is dependent on prior measures of both outcome and covariates. When the probability of attrition is dependant on unobserved measures of outcome and covariates, it is then classified as nonignorable (Laird 1988). Lastly, terminology used to illustrate the missingness process that corresponds to that of Little and Rubin (1987) was proposed by Diggle and Kenward (1994). In this terminology, completely random dropout alludes to the case of independence between assessments and attrition. Random dropout implies that dropout is a function of observed measures while informative dropout refers to dropout that depends on unobserved measures.

6.4 Methods for coping with missing data

The majority of traditional methods or strategies assume that data are ignorable i.e. either MCAR or MAR (Croninger & Douglas 2005). However, if other variables provide sufficient information about missingness to estimate an underlying relationship, then some strategies work well even if data are non-ignorable (MNAR). The problem most times is that while MCAR, MAR and MNAR provide useful distinctions of the types of missing data, it is extremely difficult, and perhaps even impossible, to determine with certainty the type of missingness obtaining in a model (Croninger & Douglas 2005). In addition, it may be possible that more than one form of missing data mechanism is present in a particular dataset (Curran et al. 1998). This therefore calls for careful consideration of the method that will be used to deal with the missing data mechanism. A distinction is therefore made between methods that assume that data are MCAR, MAR or MNAR.

6.4.1 Methods for coping with data that are MCAR

If data are MCAR, then complete case analysis can be performed. Though MCAR is the least likely assumption to be valid and difficult to verify in most studies, it is often the most assumed (Allison 2000; Schafer 1999; Schafer 1997). There are two types of complete-case analyses that can be considered under this approach, namely pairwise and listwise (casewise) deletion (Raymond & Roberts 1987). In pairwise deletion, cases which do not have data on a variable used in the current calculation are the only ones omitted. This method is desirable when the overall sample size is small or the number of cases with missing data is large. Under listwise deletion, all cases which do not have data on any of the variables used in the analysis are omitted. It is preferred over pairwise deletion when sample size is large in relation to the number of cases

which have missing data. The advantages of these two approaches include the ease and quickness with which they can be implemented, the possibility for standard complete data statistical analyses to be applied without modifications as well as the simplicity with which comparisons with other univariate statistics can be made. However, these methods produce inefficient estimates because of the inefficient use of the available information (Schafer 1997). This also leads to low power and type II errors. Biased estimates of standard errors will also result in incorrect p-values and biased parameter estimates. And unless data are truly MCAR, these methods may also lead to invalid inferences (Schafer 1999; Schafer 1997).

6.4.2 Methods for coping with data that are MAR

The majority of types of missingness that confront researchers are MAR (Foster & Bickman 2006). Therefore, a number of methods for dealing with missing data also assume that data are MAR. Popular methods when this assumption is made are listwise deletion, imputation, and analysis of incomplete data (e.g. raw maximum likelihood methods).

Listwise deletion may yield reasonable estimates even when MCAR data mechanism is relaxed to MAR. However, correct model specification would have to be used to successfully address potential bias in the restricted model (Croninger & Douglas 2005).

Imputation methods can be divided into two main groups - crude and model-based methods. Unlike the former, the latter also incorporate uncertainty around the imputed data. (Chen & Shao 2000; Schafer 1997; Rubin 1987; Little & Rubin 1987, 2002;

Cook et al. 2004; Little 1992; Raboud et al. 1996; Raghunathan et al. 2001; Skinner & Rao 2002). Six crude imputation methods and three model-based methods are considered here. The crude methods are: regression mean imputation, simple mean imputation, creating an extra category, last observation carried forward (LOCF), worst observation carried forward (WOCF) and hot-deck imputation. The model-based methods are raw maximum likelihood, expectation maximisation and multiple imputation.

In regression mean imputation (Raboud et al. 1996; Raghunathan et al. 2001; Vach & Blettner 1999), values of the variable with incomplete data are predicted from values of the variable(s) with complete data using an appropriate regression model. The predicted mean is then substituted for each unit with a missing value and information from the joint distribution of the variables is used to make the imputation. Two variations of this method can be utilised: a) stochastic mean regression imputation incorporates a random term in the prediction which is normally the regression residual from a randomly selected case with no missing values and b) pure regression imputation which does not incorporate any random term. The advantage of regression mean imputation is that it incorporates information about the covariance structure of the variables. A major disadvantage is that the datasets created may have low variability (because cases with the same values on some independent variables will all have the same value imputed for the missing observations). Another problem lies in deciding on which independent variables to include in the regression model (Raboud et al. 1996).

Under simple mean imputation, the missing values are replaced with the arithmetic mean of the observed data for that variable (Barnard & Meng 1999; Little 1992). This approach is only appropriate for continuous data. Though it is a simple and easy method to implement, it suffers from fact that it does not produce accurate estimates of measures of association or regression coefficients and variances will normally be underestimated.

Creating an extra category is used when one is dealing with categorical variables that have missing data and involves adding an extra 'missing value' category to a dataset (Baker & Laird 1988; Forster & Smith 1998; Molenberghs et al. 1999). While this again is a convenient and easy way of dealing with missingness, it has the disadvantage of having the possibility of very dissimilar classes being lumped together into one group and there is also a likelihood of severe bias in any direction arising.

The LOCF is an approach specific to longitudinal data problems. LOCF assumes that there would have been no change following attrition and therefore for each subject or individual, missing values are replaced by the last observed value of that variable. Once the data is complete, it is then analysed as if it was fully observable. LOCF is so widespread but is not underpinned by any statistical theory. In addition, means and covariance structure are seriously distorted for full longitudinal analyses. For single time point analyses the means are still likely to be distorted, measures of precision would be wrong and hence inferences would also be wrong. This is true even if the mechanism that causes the data to be missing is completely random. (Cook et al. 2004; Heyting et al. 1992; Raboud et al. 1996; Shao & Zhong 2003) .

A variant of the LOCF is the WOCF. The assumption is that the values of variables that have missing data will just get ‘worse’ (either higher or lower) and so missing values are replaced by the ‘worst’ values of that variable (Hollis 2002). After imputation, the data here is also analysed as if it was fully observed. Neither LOCF nor WOCF approaches integrate the uncertainty surrounding the imputed data in the analyses. However the methods based on statistical theory which we consider next do so (Leon et al. 2006).

The hot deck imputation method entails that a researcher makes a random draw from some appropriately chosen distribution and then consistent estimators can be obtained. Implicitly, one uses non-parametric estimates of the distribution of the missing data and this method works well with very large samples. The term ‘hot’ is used to indicate that information to fill in the missing values in the incomplete dataset is drawn from complete records of the same dataset (Skinner & Rao 2002; Vach & Blettner 1998).

The expectation maximization (EM) method is an iterative procedure that proceeds in two separate stages. During the first stage (the expectation or E stage), the expected value of the complete data log-likelihood is computed. During the maximization (M) stage, the expected values are substituted for the missing data obtained from the E stage and then the likelihood function is maximised as if no data were missing to obtain new parameter estimates. The procedure iterates through these two stages until convergence is obtained (Little & Rubin 1987).

Raw maximum likelihood methods (also called Full Information Maximum Likelihood (FIML) approaches) utilise all available data points in a database,

including the means and variances, to construct the best possible first and second order moment estimates under the MAR assumption. It can be typically represented as a covariance matrix of the variables and a vector of means (Wothke 1998). In addition to having the advantage of convenience or ease of use, it also has known statistical properties and allows for the direct computation of appropriate standard errors and test statistics as well. Its disadvantages include the difficulty to incorporate new variables to improve the accuracy of the parameter estimates of the missing values which can not, however, be utilized in the final statistical model as predictors or outcomes. It also requires specialized programming and can be very time intensive especially in large or complex models (Wothke 1998).

Multiple imputation methods make the assumption of either MAR or covariate-dependent MCAR, depending on the components of the imputation model (Allison 2000; Barnard & Meng 1999; Schafer 1997; Rubin 1987; Little & Rubin 1987, 2002). The focus of the discussion here, however, is on multiple imputation that assumes that data are MAR. Multiple imputation is a simulation-based statistical approach for the analysis of incomplete data and incorporates the uncertainty surrounding the imputed data. The pattern of missing data is traceable or predictable from other variables in the database. Application of the technique requires three steps: *imputation*, *analysis* and *pooling*. *Imputation* entails replacing each missing value with a set of $m > 1$ credible values obtained from their predictive distribution. The variation among the m imputations reflects the uncertainty with which the missing values can be predicted from the observed ones. After performing the imputations, m complete datasets are obtained. In the *analysis* of the data, any ‘complete-data’ method can then be employed. Identical analyses are carried out on each of the m datasets using any

statistical package and then Rubin's (1987) rules can be used in combining (*pooling*) the results (estimates and standard errors). Overall, estimates and standard errors that are calculated will reflect missing-data uncertainty. According to Schafer and Olsen (1998), multiple imputation is appealing for a number of reasons. First, it works in conjunction with standard complete-data methods and statistical software. In addition, the same set of m imputations is amenable to an array of analyses without the need to re-impute. The inferences (standard errors, p-values, etc.) obtained from multiple imputation are generally valid because they incorporate uncertainty due to missing data. The fact that there is high efficiency even for small values of m is another attraction. Representing the maximum fraction of missing observations by λ , the relative efficiency of an estimate based on m imputations compared to one based on an infinite number can be approximated by $(1+\lambda/m)^{-1}$ (Schafer 1997).

The propensity adjustment strategy is by and large applied to adjust for non-equivalent comparison groups in observational studies (Rosenbaum & Rubin 1983; Rosenbaum 2002; Rubin 1987; D'Agostino & Rubin 2000). A propensity score is estimated for each subject using a logit or probit regression model and is just a probability which predicts the likelihood that an observation will be assigned to a particular group, conditional on the values of the observed covariates. Therefore the missing covariates are assumed to be MAR (Rosenbaum & Rubin 1984; Rubin 1976). The propensity score summarises all the background covariates for each observation into a single index (the propensity score) allowing similar groups to be compared against each other. This propensity adjustment will only remove the bias related with the variables in the model, but will not remove bias correlated to unobserved or unmeasured characteristics. In other words, both observed and unobserved variables

have a propensity to be balanced across groups created by randomization. It is also called adjustment using propensity score derived weights (D'Agostino & Rubin 2000; Rubin 1997 and Leon et al. 2006)

6.4.3 Methods for data that are MNAR

When data are MNAR, there are more problems posed to a researcher because of the threat to a study's external validity without a succinct mechanism for dealing with potential bias (Croninger & Douglas 2005). Methods that are tenable when data are MNAR involve efforts to relax the MNAR assumption. Approaches that assume that data are MNAR tend to have a common approach which involves the joint probability modelling of both the data and the missingness mechanism (Schafer 1997). Examples of these methods are the panel selection models (Little & Rubin 1987; Schafer & Olsen 1998; Allison 2000; Heckman 1976; Amemiya 1984) and pattern-mixture approaches (Hedeker & Gibbons 1997).

The most commonly used panel selection model is the Heckman sample selection approach (Heckman 1976; 1979). It has been used extensively in econometrics where one first models the probability of an observation being observed before modelling the variable of interest. It is most effective when there is nonignorable missingness in the data.

Pattern mixture models involve categorising the different patterns of missing values in a dataset into a predictor variable. This predictor variable is then incorporated into the statistical model of interest. After the patterns of missing data are considered, the parameter estimates are then derived by pooling the stratum-specific results. The goal

is that the researcher tests for significance of the missing data patterns i.e. to determine whether or not the missing data pattern has a predictive power in the model, either by itself (a main effect) or in conjunction with another predictor (an interaction effect). These models however require a lot of programming to be able to obtain the pattern-mixture averaged results. There is also the potential of interpretation difficulties if the patterns are significant (Little 1992, 1995; Hedeker & Gibbons 1997; Little & Schenker 1995; Glynn et al. 1986).

6.5 Determining the mechanism of missingness

It is a recognised fact that data often provide little or no information at all to help determine the correct mechanism behind missingness (Heitjan 1997; Rubin 1987). In many scenarios, therefore, it is difficult, or even impossible, to know what mechanism is responsible for the missingness. Sometimes more than one mechanism may be responsible for different sets of missing data within the same evaluation (Croninger & Douglas 2005; Curran et al. 1998). This therefore means that choosing among these alternative methods is not an easy task.

There are two complementary broad approaches that can help determine the mechanism responsible for the missingness (Curran et al. 1998):

- a) Gathering information on why the missing data were not collected or obtained.
- b) Hypothesis testing of the missing data mechanism.

The first approach involves prospectively collecting as much information as possible regarding why data are missing. Questions would therefore be asked during the data

collection process about why certain data were not provided or collected. Information on the reasons why data are missing may aid in deciding whether ignoring the missing data will lead to bias in the analysis. Provided enough data has been collected, this approach will give some fairly credible indication about whether data are MCAR, MAR or MNAR (Curran et al. 1998).

The second approach involves hypothesis testing of the missing data process to determine which mechanism is behind the missingness. Several methods under this approach help the analyst to rule out that missing data are MCAR (Cohen & Cohen 1983; Orme & Reis 1991). However, there is no way of confirming that data are MCAR. Three examples of these methods are provided. The first of these methods involves creating a missing data dummy variable coded 1 if a value for variable is missing and 0 if it is not. Significant correlation between this dummy variable and other variables would imply that data are not MCAR (Acock 1997; Cohen & Cohen 1983). However, sometimes significant correlation may be as result of a larger sample size and so care must be taken when interpreting the results of this method (Musil et al. 2002). A second method is a variation of the first and requires computing statistical tests of independence between responders and non-responders after creating the missing data dummy variable. If there is significant difference between the two groups on any variables or items in the dataset, then the assumption that data are MCAR can be discounted (Fox-Wasylyshyn & El-Masri 2005; Acock 1997). Another method involves running a predictive logistic regression model where the missing data dummy variable is entered as the dependent variable into the model. Any significant association between the dependent variable and the independent variables rules out the possibility that data are MCAR (Miller & Wright 1995; Hair et al. 1998;

Little & Rubin 1987). The method by Ridout (1997) extends this methodology by considering missing data at more than two different time points for longitudinal data. In this method, the missing data dummy variable can be entered as an independent variable together with other variables in a regression model. The dependent variable can be any outcome of interest. This specification ensures that the effect of complete data on the relationship between the missing data dummy variable and the dependent variable is accounted for. If the dummy variable is significantly related to the dependent variable, then the missing data can not be said to be MCAR (Orme & Reis 1991).

If information about the missing or nonrespondent data is not available, it will be impossible to ascertain if data are MAR or MNAR (Little & Rubin 1987). Testing for these two mechanisms will always rely on strong but often untestable assumptions (Molenberghs et al. 1999; Potthoff et al. 2006). The default position for most statistical software is to assume that missing data are MCAR. Although this assumption may hold in certain situations, it is a strong assumption, which is usually violated in many applications (Durrant 2005). The MAR assumption is a relaxation of the MCAR condition and is also the most assumed mechanism of missingness for most analyses (Potthoff et al. 2006). While some argue that in many instances, results obtained using MAR-based analyses are valid when departures from MAR are not big enough (Schafer et al. 2002) others submit that applying MAR-based methods may leads to seriously biased results when the actual missing data mechanism is MNAR (Greenlees et al. 1982; Collins et al. 2001). In the absence of extra information about the reasons that have led to missing data, care should be taken that selection of methods for dealing with the missing data is not random.

6.6 Key messages for health services researchers from this chapter

The fundamental message from this chapter is that ignoring missing data may lead to biased results and therefore misleading conclusions. Methods for dealing with missing data are premised on the mechanism responsible for the missingness. There are three major mechanisms that may explain why data are missing: MCAR, MAR and MNAR. MCAR is when the probability that an item is missing is unrelated to any measured or unmeasured characteristic for that unit. Complete case analysis (listwise and pairwise deletion) can be used for coping with missing data when missing data are assumed to be MCAR. Missing data are said to be MAR if the missingness is related to another observed variable. Methods for dealing with missing data when these data are assumed to be MAR include listwise deletion, various imputation methods, expectation maximisation and raw maximum likelihood approaches. MNAR is when the probability of missingness depends on the values of both the missing and observed values as well as on the values of one or more other variables in the analytic model. Missing data methods that assume that data are MNAR include those based on panel selection and pattern mixture models. To determine the mechanism that is responsible for missingness, two broad but complementary approaches can be used: prospectively gathering information on why the missing data were not collected or obtained and hypothesis testing of the missing data mechanism. However, it is not always possible to ascertain the missing data mechanism(s) but where possible, it is advisable to establish this mechanism(s).

6.7 Conclusion

This chapter has considered three different mechanisms that may be responsible for missing data and then discussed approaches that can be used to deal with the missing

data. These approaches are linked to these mechanisms. The hypothesis-based techniques for detecting the pattern of missingness are limited in that they can only be used to rule out MCAR but can not confirm this mechanism. Further, there are no hypothesis-test-based techniques available for determining if data are MAR or MNAR. This therefore means that unless there is extra information gathered during the data collection exercise about the cause of missingness, there should not be any arbitrary selection of assumptions behind data missing mechanisms.

The next chapter tests the methods of dealing with missing data on the demonstration dataset where extra information was available that gave strong indications of the reasons for missing data in the dataset (and therefore the missingness mechanism). The analysis tests whether different results can be obtained from methods that assume a different mechanisms from that that could be deduced from the extra information obtained from the demonstration dataset.

CHAPTER SEVEN – CHALLENGES WITH MISSING DATA: EMPIRICAL ANALYSIS

7.1 Introduction

This chapter is empirical and seeks to examine the effect of using different approaches for coping with missing data. Each method assumes a different mechanism behind the missingness. The three mechanisms that can be responsible for missing data were described in chapter six. This analysis tests what happens when extra information that suggests that a particular mechanism is responsible for missing data is disregarded and methods for dealing with the missing data are chosen arbitrarily. Using a regression framework, the factors that explain variation in costs per patient, change in EQ-5D (Δ EQ-5D) and change in the Barthel index (Δ Barthel) of intermediate care patients were explored. Three methods were used: complete case analysis (assuming MCAR), multiple imputation (assuming MAR) and Heckman selection model (assuming MNAR). The chapter also considers the implications for health services researchers when using these methods. The demonstration dataset was used in all these analyses.

7.2 Missing data in the demonstration dataset

The variables that were collected in the demonstration dataset have been previously presented in Table 4.2 of chapter four. Up to 42% of the data were missing for some variables in that dataset. For purposes of comparing the methods described in section 7.3, a decision was made to focus on missingness only in the dependent variables i.e. missingness in the cost per patient, Δ EQ-5D and Δ Barthel. As a result, there were no missing data for any of the independent variables. A sample of 717 individuals was used for the cost per patient models and 125 (17.4%) of these individuals had missing

observations on the cost per patient variable. For the Δ EQ-5D and Δ Barthel models, a sample of 1105 individuals was utilised. Of this sample, 417 (37.7%) and 392 (35.5%) of the total sample had missing values on the Δ EQ-5D and Δ Barthel variables, respectively.

7.3 Reasons for missing data in the demonstration dataset

The cost per patient variable was calculated by combining resource data with budget information for the individual intermediate care services. However, because of time constraints placed on the data collection process, it was not possible to collect all the cost data. No other reason was established as being responsible for the missing cost data. This suggests that where cost data were missing, it would be reasonable to assume that these data were MCAR

Information obtained from the intermediate care coordinators about the missing EQ-5D and Barthel data indicated that some services did not routinely collect this information (ICNET 2005). This suggested that it was plausible to assume that the missingness mechanism for these missing data was MCAR. However, another reason established was that some patients were not able to self-report some of these data because of their functional status. This is borne out for the EQ-5D through examining the mean Barthel scores for individuals who had missing values on the EQ-5D. Where data were available, the Barthel scores for individuals who had missing EQ-5D scores were on average lower than those for individuals who did not have missing EQ-5D information. Lower functional status could therefore be linked to missing EQ-5D scores. This may suggest that that some of these missing EQ-5D data were MAR.

As will be shown in chapter nine, there is a positive relationship between EQ-5D and Barthel scores. This therefore means that it was possible that the probability of having missing information on the EQ-5D (a measure of health status) was linked to an individual's actual health status i.e. the poorer ones' health status was, the more difficult it was for them to provide data on the EQ-5D. It was therefore reasonable to assume that some of the missing data on the EQ-5D could also have been MNAR.

As shown in chapter five, the Δ EQ-5D and Δ Barthel scores were calculated by subtracting the scores at admission from those at discharge. A number of individuals had however been transferred to other services before the end of their intermediate care episode. For some of these, it meant that their EQ-5D and Barthel scores at 'discharge' were not collected making it impossible to compute the Δ EQ-5D and Δ Barthel variables. This again could be seen as a situation where the data were MAR as the reason for the patients transfer was sometimes linked to their health or functional status e.g. the more functionally independent an individual was, the more likely they were to be transferred to a less intensive form of intermediate care.

From this extra information collected, it was reasonable to assume that missing cost data were MCAR. On the other hand, MCAR, MAR and MNAR could all be assumed as the reason for the missing data on the EQ-5D and Barthel.

7.4 Approaches for dealing with the missing data

Three methods, each assuming either MCAR, MAR or MNAR, were used on the two samples obtained from the demonstration dataset and described in section 7.2. A

regression framework was used and in general, the regression relationship between the outcomes of interest and the independent variables could be illustrated as:

$$Y_i = f(X_i) + \mu_i \quad (7.1)$$

where Y_i denotes the outcome of interest (cost per patient, Δ EQ-5D or Δ Barthel) for the i th individual,

X_i is a vector of explanatory variables for the i th individual and

μ_i is the stochastic error term for the i th individual.

A total of six sets of regression models (two for each method) were conducted:

Method 1 (assuming data are MCAR) – Regression on complete cases

- a. A generalised linear regression model (GLM) used on complete cases (dependent variable is cost per patient). Algebraically, this model could be depicted as:

$$C_i = f(X_i) + \mu_i \quad (7.2)$$

$$E(C/X_i) = g^{-1} f(X_i) \quad (7.3)$$

where:

$E(C/X_i)$ = is the conditional mean of cost for patient i given X_i ,

g = monotonic differentiable link function describing how $E(C/X_i)$ is

related to $f(X_i)$ and

X_i and μ_i are as described in equation (7.1)

- b. Ordinary least squares (OLS) regression models run on complete cases (dependent variables were $\Delta\text{EQ-5D}$ and $\Delta\text{Barthel}$). Algebraically:

$$O_i = f(X_i) + \mu_i \quad (7.4)$$

$$E(O/X_i) = f(X_i) \quad (7.5)$$

where:

$E(O/X_i)$ = is the conditional mean of $\Delta\text{EQ-5D}_i$ or $\Delta\text{Barthel}_i$ given X_i and

X_i and μ_i are as described in equation (7.1)

Method 2 (assuming data are MAR) – Regression on multiply imputed datasets⁵

- c. GLM administered on multiply-imputed (MI) datasets using costs per patient as the dependent variable.
- d. OLS regression models run on MI datasets using $\Delta\text{EQ-5D}$ and $\Delta\text{Barthel}$ as dependent variables

⁵ Multiply imputing missing values for variables that will be used as dependent variables in subsequent regression analyses has been shown to be an acceptable practice (Schafer & Olsen 1998).

Method 3 (assuming data are MNAR) – Heckman selection models

- e. Heckman selection regression model with the log of cost per patient as the dependent variable. The Heckman model is made up of two equations: the regression model and the selection model.

The regression model could be depicted as follows:

$$c = x\beta + \mu_1 \quad (7.6)$$

For the selection model,

$$c \text{ is observed if } z\gamma + \mu_2 > 0 \quad (7.7)$$

where:

c = cost per patient; x = independent variables in the regression model;

μ_1 = error term for the regression model; β = coefficient for x ;

z = independent variables for the selection model; γ = coefficient for z ;

μ_2 = error term for the selection model;

and in addition the following hold:

$\mu_1 \sim N(0, \delta)$ i.e. μ_1 is normally distributed with mean 0 and variance δ ;

$\mu_2 \sim N(0, 1)$ i.e. μ_2 is normally distributed with mean 0 and variance 1;

$\text{corr}(\mu_1, \mu_2) = \rho$ i.e. μ_1 and μ_2 are correlated with correlation equal to ρ .

When $\rho = 0$, then OLS regression provides unbiased estimates while $\rho \neq 0$ results in biased estimates.

- f. Heckman selection regression model using $\Delta EQ-5D$ and $\Delta Barthel$ as dependent variables.

The regression model could be depicted as follows:

$$o = x\beta + \mu_1 \quad (7.8)$$

For the selection model,

$$o \text{ is observed if } z\gamma + \mu_2 > 0 \quad (7.9)$$

where:

$o = \Delta EQ-5D$ or $\Delta Barthel$ and $x, \mu_1, \beta, z, \gamma$ and μ_2 are as described in equations (7.6) and (7.7). The assumptions made in these two equations also hold here.

Method 1 uses approaches that assume that missing data are MCAR. Complete case regression using GLM or OLS models was conducted. Cases which did not have data on any variable used in the current equation only were omitted (pairwise deletion). As a result, samples of 592, 688 and 713 observations were used for the cost per patient, $\Delta EQ-5D$ and $\Delta Barthel$ models, respectively. These results are therefore only for subjects with no missing data in any particular equation. This type of analysis is the default in most statistical packages.

In method 2, the MAR assumption was posited for the missing data. Up to about 38% of the data were missing and multiple imputation datasets were created to account for

these missing data before running GLM and OLS regression models. Multiple imputation assumes that the data is MAR, that the likelihood of missing data on any variable is not related to its particular value. These analyses focussed on imputing values for the dependent variables where the independent variables were not missing. This imputation created complete datasets where there was no missing data i.e. 717 observations for the cost per patient model and 1105 observations for the Δ EQ-5D and Δ Barthel models. The rationale for this particular imputation was to allow for direct comparison between the results obtained using this method and those produced by method 3, which comparison required that the same samples were analysed. Five sets of imputations were created. Since there was up to 38% data missing, these imputations led to point estimates that were at least $(1+0.38/5)^{-1} = 93\%$ as efficient as those based on $m = \infty$ imputations (Schafer 1997).

Method 3 assumed that missing data on the dependent variables were MNAR and a Heckman sample selection regression model was used on complete cases. Whereas method 1 only considered cases where there was no missing data for both the dependent variable and independent variables, method 3 considers all subjects including those that had missing cost, EQ-5D or Barthel information. The sample selection used a dummy variable equal to 1 if the dependent variable was not missing and equal to 0 if it was. Using this classification, 125 out of 717 observations were censored (missing) for the cost per patient model. And out of 1105 observations, 417 and 392 were censored for the Δ EQ-5D and Δ Barthel models, respectively. The sample selection results are shown in Table A5 in the appendix. The instrumental variable used was the dummy equal to 1 if a service was performing a supported discharge function and equal to 0 otherwise. The justification for this choice was the

fact this dummy was previously not significant in explaining variation in the costs per patient, Δ EQ-5D or Δ Barthel scores in the complete cases regression models (method 1). The fact that this dummy was later significant in the sample-selection model (Table A5 in the appendix) seems to suggest its suitability as an instrumental variable. The interpretation of this result would be that intermediate care services that were performing a supported discharge function were more likely to have costs or outcome information (Δ EQ-5D or Δ Barthel) collected than those that did not.

7.4.1 Choice of regression families

In this exercise, it was important to compare both the signs and sizes of coefficients from the different regression models. In chapter four and chapter five, both costs per patient and outcome variables (Δ EQ-5D and Δ Barthel) were shown to be skewed. All three dependent variables were also found to be heteroscedastic in their residuals. The GLM was shown to be a principled approach for dealing with the twin problems of skewness and heteroscedasticity in chapter five. Another method for dealing with skewed data is the use of log-transformation where the natural log of the dependent variable is obtained (Altman 1991) despite the limitations outlined in section 5.3.3 of the same chapter. The exponentiated coefficients from the GLM model have been shown to be easily comparable to the exponentiated counterparts obtained from a log-transformed model (Kilian et al. 2002). For the cost models, therefore, a decision was made for the GLM to be used for both the complete cases and the multiply imputed datasets while a log transformed cost per patient was used in the Heckman regression model. A different approach was taken for the outcome dependent variables (Δ EQ-5D and Δ Barthel). This was because these variables also had negative values. As a result, log transformation of these variables would have required the use of a shift factor and

the transformed variables would then have had to be appropriately retransformed once the results of the model had been obtained. However for ease of analysis and comparison, a decision was made to use the raw scale of these variables. As a result, OLS regressions were used for both the Δ EQ-5D and Δ Barthel in the regression on complete cases and on multiply imputed datasets. The raw scale of the two variables was also used in the Heckman selection models. Multiple imputations were conducted in NORM (Schafer 1999) while the rest of the analyses were done in STATA version 8.2 (StataCorp LP 2004).

7.5 Results

The results of the above analyses are presented in Tables 7.1 and 7.2 for the costs per patient models and Tables 7.4 to 7.7 for the Δ EQ-5D and Δ Barthel models.

Cost per patient models

The results of the GLM regression model on complete cases (method 1) and GLM regression model on multiply imputed datasets (method 2) are similar. As shown in Table 7.1, all of the variables that were found to be significant in method (2) were also significant in method (1) with the exception of one (acute admission avoidance service) which was significant in model (2) only. Also, the size of coefficients for nearly all of these variables differed by less than 3.4% except the one for ‘completed IC episode’ which differed by about 14.4%. The sizes of the standard errors were also similar. Further, the variables significant in both models had the same direction of influence on costs per patient (Table 7.2). On the other hand, the results obtained from the Heckman selection regression model (method 3) were much more different. A lot more variables were found to be insignificant with only two variables shown to

significantly influence costs per patient. The sizes of the coefficients in the Heckman model were also different from those of the other two methods. For instance, the coefficient for ‘acute admission avoidance service’ was about 730 times bigger than that in obtained in method (2). The mills ratios were -3.402 and -4.506 for the Heckman selection models with and without interactions, respectively. These were both statistically significant at 95% level of significance.

Change in EQ-5D models

Here, the results from all three models/methods were broadly similar. Nearly all of the variables that were significant in one model were also significant in the other models. The only exception were the ‘duration of service provision’ and ‘Alternative to IC-Other*Type of IC’ (both only significant in method 2), ‘acute admission avoidance service’ (only significant in method 3) and ‘alternative to IC-Other’ (significant only in models 1 and 3). The sizes of the coefficients of variables commonly significant in all models differed at most by about 22% with the standard errors differing at most by 42% (Table 7.3). Further, the variables significant in all three models had the same direction of influence on the change in EQ-5D (Table 7.4). The mills ratios were -0.284 and -0.143 for the Heckman selection models with and without interactions, respectively. These were both statistically significant at 95% level of significance.

Change in Barthel models

As in the ‘change in EQ-5D models, the results obtained from all three models/methods for the change in Barthel were broadly similar (Table 7.5). All of the variables that were significant in one model were also significant in the other models with the exception of ‘acute admission avoidance service’ (only significant in method

3) and 'Other IC Outcome' variable only significant in both method (1) and method (2). However, the differences in terms of the sizes of coefficients and standard errors of variables significant in all methods were slightly bigger in these models than in the 'change in EQ-5D models'. They differed at most by about 54% and 322% for coefficients and standard errors, respectively. Table 7.6 shows that the variables significant in all three models had the same direction of influence on the change in Barthel. The mills ratios were -1.662 and -0.101 for the Heckman selection models with and without interactions, respectively. These were both statistically significant at 95% level of significance

7.6 Discussion

7.6.1 Key findings

Why consider methods for dealing with missing data

As was shown in Table 4.2 of chapter four, the demonstration dataset had a significant amount of missing data. Up to 42% and 38% of the data on EQ-5D and Barthel scores, respectively, were missing while 31% of the sample had missing cost data. Further, all but one variable in the dataset (Type of IC) had missing data ranging from 3 to 18%. This situation is not exclusive to this dataset as was shown in chapter three (section 3.5.2.1). If these missing data are simply ignored, then there is a chance that biased and underpowered results may be obtained (Roderick et al. 2001; Schafer, 1997). The most appropriate method of dealing with this amount of missingness therefore had to be determined (Cohen & Cohen 1983; Curran et al. 1998).

What was shown

The evidence gathered concerning the missing cost data strongly suggested MCAR as the reason for this missingness. This implied that the results obtained from an approach that assumed this missingness mechanism would be reasonably credible. When MAR was assumed, the results obtained were not significantly different from those based on the MCAR assumption. These findings seem to bear out the position held by Schafer et al (2002) and David et al (1986) that in many realistic applications, departures from MAR are not big enough to effectively invalidate the results of an MAR-based analysis. However, care must be taken not to apply MNAR-based methods when the missing data are clearly not MNAR as MNAR approaches often require assumptions that cannot be validated from the data at hand (Verbeke & Molenberghs, 2000). MNAR-based approaches are best implemented as sensitivity analyses so as to assess how robust results are across different analytic approaches (Mallinckrodt et al. 2003). The use of an MNAR-based method in the costs per patient model yielded results that were so different to those obtained when either MCAR or MAR were assumed. In particular, fewer significant variables were obtained in the MNAR-based method while the sizes of the coefficients were larger. Different conclusions would therefore be reached if the MNAR assumption was made for the missing cost data.

All three mechanisms of missingness were shown to be potential causes of the missing EQ-5D and Barthel data. The results from the Δ EQ-5D and Δ Barthel models show that the choice of mechanism did not have any significant effect on the results. Despite the sizes of the coefficients and standard errors being somewhat different, the results from all three methods were broadly similar and therefore similar conclusions

could have been reached. A possible explanation for this may have been the fact that the reason for missing data could be ascribed to any one of the three mechanisms of missingness or indeed a combination of these mechanisms. While the extra information gathered during the data collection process supported the assertion that the missing data were either MCAR, MAR or MNAR, the significant mills ratios lent further support to the MNAR assumption as its significance in the selection models indicated the presence of significant selection bias. However, selection models, even though identifiable, should be treated with caution especially when data are not MNAR (Glynn et al. 1986). Pattern mixture models would be another alternative (Hedeker & Gibbons 1997; Hogan & Laird 1997).

Why does the method for accounting for missing data matter?

What these results from all six regression models showed was that when missing data were MCAR, the use of the MAR assumption did not change the results significantly. However, the use of MNAR-based methods did. These results therefore show that it is important to be certain about the missingness mechanism as this choice is not inconsequential. If there was a possibility that all of the three missingness mechanisms were responsible for the missing data, the results obtained showed that similar results would be obtained regardless of the method used. However, this result should be viewed cautiously as it may just reflect some peculiarity with this data. One of these peculiarities may be the fairly large size of the dataset which Croninger and Douglas (2005) assert leads to fairly comparable results regardless of the method used. The results of these analyses would also need to be validated on other dataset before generalising these findings.

Many studies have emphasised the importance of determining the mechanism behind missing data before deciding on the technique to use (Cohen & Cohen 1983; Orme & Reis 1991; Curran et al. 1998). These results also seem to underline the fact that, where possible, efforts should be made to ascertain the reason that has led to missing data and then determine the method of analysis to be used. This therefore seems to strongly suggest that erroneous results would be obtained if the choice of the techniques to be used for dealing with missing data (based on assumptions about the missingness mechanisms) is made arbitrarily (Little & Rubin 1987; Molenberghs et al. 1999; Potthoff et al. 2006). According to Curran et al (1998), the problem of missing data is best solved by making attempts not to have any missing data in the first instance. But where missing data are encountered, then sound statistical methods of analysing missing data based on the missingness mechanism must be employed.

7.6.2 Key messages for health services researchers from this chapter

Extra information collected prospectively about why data are missing can help inform the choice of the method to be used for dealing with the missingness. When MCAR and MAR-based methods were used for data that was in reality was MCAR, similar results were obtained. However, using a MNAR-based method on these data yielded substantially different results. The key message from these analyses therefore seems to be that MAR-based methods are robust to departures from the MAR and can therefore be used to produce valid results even when missing data are actually MCAR. Caution however must be placed on using MNAR-based methods when missing data are actually MNAR.

7.7 Conclusion

The extra information gathered about the reason for the missingness suggested that the missing data on the cost per patient variable were MCAR while that on the outcome variables were MCAR, MAR or MNAR. The results obtained from the cost model showed that assuming MCAR or MAR resulted in similar results but assuming MNAR yielded substantially different findings. The results from the Δ EQ-5D and Δ Barthel models were not substantially different but the sizes of coefficients and standard errors were disparate. The assumption made about the missing data is therefore not without consequence. Presuming either MCAR or MNAR requires stronger assumptions than MAR and therefore one needs to be very certain about the missingness mechanism before choosing any of these assumptions.

The next two chapters consider the last of the statistical problems being considered in this thesis, which is the issue of predicting utility measures of outcome from non-utility ones.

Table 7.1: Comparison of Results from three methods of Regression Analysis of costs per patient

		<i>GLM on complete cases n = 592 [1]</i>		<i>GLM on MI dataset n = 717 [2]</i>		<i>Heckman on complete cases n = 717, 125 obs censored [3]</i>	
Variables		Exp (Coeff)	S.E.	Exp (Coeff)	S.E.	Exp (Coeff)	S.E.
Episode Characteristics	Age in 2003	0.996	0.003	0.997	0.002	1.000	0.012
	Gender	0.982	0.063	1.009	0.060	1.085	0.281
	Lives alone	1.052	0.059	1.047	0.056	1.106	0.275
	Barthel score at admission	0.973	0.009**	0.984	0.008*	0.884	0.061*
	EQ5D score at admission	0.973	0.090	0.935	0.087	1.400	0.435
Descriptors of IC Service	Acute Admission Avoidance Service	0.930	0.129	0.812	0.092*	6.723	0.960*
	Type of IC	3.181	0.079**	3.150	0.070**	5.146	1.274
	Transferred before end of IC episode	1.144	0.310	1.259	0.258	1.316	1.422
	Completed IC episode	2.094	0.300*	2.396	0.248**	4.611	1.318
	Other IC Outcome	2.703	0.337**	2.796	0.287**	4.374	1.475
	Patient Died (Reference. Group)						
Descriptors of IC-related Services	Referral – Primary	0.777	0.123*	0.764	0.121*	0.936	0.576
	Referral – Hospital	0.914	0.158	0.777	0.134	4.523	0.930
	Referral – Other	1.001	0.212	0.935	0.195	2.240	0.984
	Referral – Social Workers (Reference Group)						
	Alternative to IC – Other	1.053	0.079	1.058	0.077	0.508	0.451
	Alternative to IC – Home	1.121	0.074	1.058	0.070	1.112	0.329
Interactions	Alternative to IC – Hospital (Reference Group)						
	Barthel score at admission*Type of IC	1.031	0.018	1.017	0.097	1.131	0.092
	Acute Admission Avoidance Service* Type of IC	1.214	0.163	1.217	0.136	0.579	0.752
	Transfer before IC end*Type of IC	1.145	0.185	1.176	0.169	1.145	0.825

Completed Episode*Type of IC	1.152	0.195	1.112	0.162	0.240	0.952
Other IC Outcome*Type of IC	0.717	0.708	0.583	0.534	0.773	2.846
Patient died*Type of IC (Reference group)						
_constant	1140.3	0.421**	951.5	0.360**	345.3	1.866**
N		592	717			717
Censored obs						125
R-Squared		0.359				0.634
Rho						0.950

* 5 % level of significance, ** 1 % level of significance; Dependent variable: cost per patient for GLM and log of cost per patient for Heckman Selection model,
IC = Intermediate care

Table 7.2: Comparison of significant variables in the three cost models **

Significant variables		<i>GLM on complete cases n = 592</i> [1]	<i>GLM on MI dataset n = 717</i> [2]	<i>Heckman on complete cases n = 717, 125 obs censored</i> [3]
Episode Characteristics	Barthel score at admission	-	-	-
Descriptors of IC Service	Acute Admission		-	+
	Avoidance Service			
	Type of IC	+	+	
	Transferred before end of IC episode			
	Completed IC episode	+	+	
	Other IC Outcome	+	+	
	Patient Died (Reference. Group)			
Descriptors of IC-related Services	Referral – Primary	-	-	
	Referral – Hospital			
	Referral – Other			
	Referral – Social Workers (Reference Group)			

** + (-) depicts a positive (negative) relationship between (log) cost per patient and the independent variable, IC = Intermediate care

**Table 7.3: Comparison of Results from three methods of Regression Analysis
(Change in EQ5D)**

		<i>OLS on complete cases n = 688</i>		<i>OLS on MI dataset n = 1105 cases</i>		<i>Heckman on complete cases n = 1105, 417 obs censored</i>	
		<i>[1]</i>		<i>[2]</i>		<i>[3]</i>	
	Variables	Coeff	S.E.	Coeff	S.E.	Coeff	S.E.
Episode Characteristics	Age in 2003	0.000	0.001	0.000	0.001	0.000	0.001
	Gender	0.046	0.022*	0.051	0.018**	0.054	0.024*
	Lives alone	0.020	0.020	0.015	0.017	0.029	0.023
	Barthel score at admission	0.017	0.003**	0.017	0.002**	0.016	0.003**
	EQ5D score at admission	-0.495	0.033**	-0.479	0.026**	-0.484	0.037**
Descriptors of IC Service	Acute Admission	-0.038	0.027	-0.017	0.021	0.156	0.042**
	Avoidance Service						
	Duration of Service Provision	0.000	0.000	0.001	0.000*	0.000	0.000
Descriptors of IC-related Services	Referral – Primary	-0.031	0.052	-0.044	0.043	-0.020	0.058
	Referral – Hospital	-0.098	0.051	-0.053	0.042	0.020	0.059
	Referral – Other	-0.003	0.078	0.059	0.065	0.013	0.086
	Referral – Social Workers (Reference Group)						
	Alternative to IC – Other	-0.063	0.031*	-0.077	0.025**	-0.077	0.030*
	Alternative to IC – Home	-0.045	0.023*	-0.028	0.019	-0.046	0.022*
	Alternative to IC – Hospital (Reference Group)						
Interactions	Gender*Type of IC	-0.048	0.053	-0.027	0.037	-0.057	0.053
	Barthel score at admission*Type of IC	0.003	0.004	-0.002	0.003	0.004	0.004
	EQ5D score at admission *Type of IC	-0.098	0.083	0.061	0.057	-0.118	0.082
	Acute Admission Avoidance Service*Type of IC	0.110	0.064	0.039	0.039	0.086	0.063
	Alternative to IC – Other *Type of IC	0.137	0.084	0.133	0.059*	0.140	0.082
	Alternative to IC – Home*Type of IC	0.086	0.106	-0.027	0.049	0.070	0.104
	Alternative to IC – Hospital *Type of IC						

(Reference Group)						
_constant	0.157	0.101	0.093	0.084	0.100	0.105
N		688	1,105			688
Censored obs						417
R-Squared		0.284	0.266			0.634
Rho						0.950

* 5 % level of significance, ** 1 % level of significance;

Dependent variable: change in EQ-5D, IC = Intermediate care

Table 7.4: Comparison of significant variables in the three EQ5D models **

Significant variables		<i>OLS on complete cases n = 688</i> [1]	<i>OLS on MI dataset n = 1105 cases</i> [2]	<i>Heckman on complete cases n = 1105, 417 obs censored</i> [3]
Episode Characteristics	Gender	+	+	+
	Barthel score at admission	+	+	+
	EQ5D score at admission	-	-	-
Descriptors of IC Service	Acute Admission			+
	Avoidance Service			
	Duration of Service Provision		+	
Descriptors of IC-related Services	Alternative to IC – Other	-	-	-
	Alternative to IC – Home	-		-
	Alternative to IC – Hospital (Reference group)			
Interactions	Alternative to IC – Other*Type of IC		+	
	Alternative to IC – Home*Type of IC			
	Alternative to IC – Hospital*Type of IC (Reference group)			

** + (-) depicts a positive (negative) relationship between change in EQ-5D and the independent variable, IC = Intermediate care

**Table 7.5: Comparison of Results from three methods of Regression Analysis
(Change in Barthel)**

		<i>OLS on complete cases n = 712</i>		<i>OLS on MI dataset n = 1105 cases</i>		<i>Heckman on complete cases n = 1105, 392 obs censored</i>	
		<i>[1]</i>		<i>[2]</i>		<i>[3]</i>	
	Variables	Coeff	S.E.	Coeff	S.E.	Coeff	S.E.
Episode	Age in 2003	-0.010	0.009	-0.009	0.007	-0.011	0.009
Characteristics	Gender	-0.007	0.208	0.097	0.164	0.037	0.218
	Lives alone	0.225	0.190	0.181	0.150	0.320	0.202
	Barthel score at admission	-0.318	0.028**	-0.325	0.022**	-0.305	0.030**
	EQ5D score at admission	-0.343	0.312	-0.428	0.239	-0.216	0.328
Descriptors of IC Service	Acute Admission	0.103	0.218	0.060	0.167	0.728	0.337*
	Avoidance Service						
	Duration of Service Provision	0.008	0.003*	0.011	0.003**	0.006	0.003*
Descriptors of IC-related Services	Transfer before IC end	4.084	2.452	0.559	0.607	2.713	2.348
	Completed Episode	7.438	2.440**	3.443	0.587**	4.926	2.478*
	Other IC Outcome	6.640	2.477**	2.921	0.656**	4.727	2.432
	Patient died (Reference group)						
	Alternative to IC – Other	-1.130	0.291**	-1.076	0.221**	-1.267	0.291**
	Alternative to IC – Home	-0.709	0.223**	-0.667	0.169**	-0.669	0.219**
	Alternative to IC – Hospital (Reference Group)						
Interactions	Barthel score at admission*Type of IC	-0.071	0.050	-0.035	0.027	-0.072	0.051
	Acute Admission Avoidance Service* Type of IC	0.592	0.575	0.131	0.354	0.599	0.589
	Duration of Service Provision*Type of IC	-0.006	0.008	0.001	0.006	-0.006	0.008
	Transfer before IC end*Type of IC	-0.299	0.979	-0.160	0.411	-0.300	0.962
	Completed Episode*Type of IC	1.053	0.830	0.374	0.424	1.055	0.816
	Other IC Outcome*Type of IC	0.189	2.000	0.072	0.889	0.200	1.980
	Patient died*Type of IC (Reference group)						

Alternative to IC – Other *Type of IC	0.796	0.793	0.968	0.543	0.795	0.778
Alternative to IC – Home*Type of IC	2.261	1.025*	0.124	0.447	2.261	1.006*
Alternative to IC – Hospital *Type of IC (Reference Group)						
_constant	0.046	2.536	3.888	0.843	2.687	2.592
N		713	1,105			713
Censored obs						392
R-Squared		0.278				0.634
Rho						0.950

* 5 % level of significance, ** 1 % level of significance;

Dependent variable: change in Barthel, IC = Intermediate care

Table 7.6: Comparison of significant variables in the three Barthel models **

Significant variables		<i>OLS on complete cases n = 688 [1]</i>	<i>OLS on MI dataset n = 1105 cases [2]</i>	<i>Heckman on complete cases n = 1105, 417 obs censored [3]</i>
Episode Characteristics	Barthel score at admission	-	-	-
Descriptors of IC Service	Acute Admission			+
	Avoidance Service			
	Duration of Service Provision	+	+	+
Descriptors of IC-related Services	Transferred before end of IC episode			
	Completed IC episode	+	+	+
	Other IC Outcome	+	+	
	Patient Died (Reference. Group)			
	Alternative to IC – Other	-	-	-
	Alternative to IC – Home	-	-	-
	Alternative to IC – Hospital (Reference group)			

** + (-) depicts a positive (negative) relationship between change in Barthel and the independent variable, IC = Intermediate care

CHAPTER EIGHT - CHALLENGES WITH OUTCOME DATA: LITERATURE REVIEW

8.1 Introduction

Health related quality of life (HRQoL) and functional independence of older people are important outcome measures for evaluating efficacy and effectiveness of therapeutic interventions (Coast et al. 1998; Brazier et al. 1996; King 1996; Griffiths et al. 2001; Yohannes et al. 1998; van den Bos & Triemstra, 1999; Davies 1996; Lyons et al. 1997; Gosman-Hedstrom & Svensson, 2000; Kwon et al. 2004; van Exel et al. 2004; Sainsbury et al. 2005; Gauggel et al. 2004). The former gives an indication of an individual's health status while the latter permits an understanding of how limited an individual is by his/her functional condition. In many well designed evaluation studies on older people, preference or utility-based health related quality of life (HRQoL) data are collected using an appropriate instrument. Such utility-based data can then be used in economic evaluations such as cost-effectiveness or cost-utility analysis to be undertaken. One such instrument is the EQ-5D whose construct validity when used on populations of older people has been well documented (Coast et al. 1998; Brazier et al. 1996; van den Bos & Triemstra 1999; Lyons et al. 1997).

When conducting evaluations of services for older people, however, it is not always possible to collect data on outcomes, such as the EQ-5D, which can be readily converted into utilities. Some older people are physically or mentally not able to provide self-report HRQoL data and so the issue is a missing data problem. Sometimes data on HRQoL are simply not available at all because they have not been collected. This tends to be particularly the case when researchers are reliant on routinely collected data for service evaluations. In such instances, it is important to

avoid the bias that might result from simply ignoring the individuals with missing data. In addition, older people may receive care from different service providers who in turn do not always use the same instruments to document HRQoL. These scenarios therefore highlight the general limitations of relying on just one instrument.

Sometimes, information on other non-utility outcome measures may have been collected. A question then arises as to whether this information can be used in any way to obtain utilities that can thereafter be used for economic evaluations. One approach is predicting or mapping utilities from the non-utility outcomes using regression methods. A method of predicting a utility-based outcome measure from a non-utility-based one will be invaluable when the ultimate goal is to carry out an economic evaluation. It is even more attractive when the non-utility-based index is disease or condition-specific and therefore more acceptable to staff working in a particular area as well more sensitive to health status differences or changes for individuals with a specific disease (Froberg & Kane 1989). This chapter reports the results of a literature review of studies where utility-based outcome measures have been predicted from non-utility based ones. The key messages for health services researchers working in this and other related areas are presented in the discussion section. The results of this review informed the empirical analysis that is conducted in chapter nine where the possibility of predicting the EQ-5D, a utility-based measure, from the Barthel index, a non-utility based measure, is examined. A distinction is first made between utility and non-utility measures.

8.2 Utility versus non-utility measures

Utility measures have also been referred to as index measures (Johnson & Coons 1998), clinimetric outcomes (Ribera et al. 2006), health values (Tsevat et al. 1996) or preference based measures (Brazier et al. 2007). Utility measures refer to the ‘cardinal values that represent the strength of an individual’s preferences for specific outcomes under conditions of uncertainty’ (Torrance 1986; Torrance & Feeney 1989). In particular, health utilities are a valuation of preferences for specified health states or treatments. The valuations are normally based on weights obtained from patients or the general public. They reveal someone’s opinion of the relative desirability for different states of health. Utility-based measures result in a single index that can compare one quality of life or health state to others and also has the property of being applicable across different healthcare fields (Velikova et al. 1999; Brazier et al. 1998; Feeny & Torrance 1989).

Many policy makers use economic evaluation to make decisions about costs and outcomes of older people (Drummond et al. 1997). Utility analysis has been found to be useful as it permits quality of life adjustments to be made to a given set of outcomes. Utilities make it possible to construct a generic outcome measure, the quality adjusted life year (QALY), which can be used in different forms of economic evaluation (Drummond et al. 1997). A number of bodies, including the US Panel on Cost-effectiveness in Health and Medicine have recommended the use of QALYs in cost-effectiveness analyses (Gold et al. 1996). The use of utility measures has been proposed as way of conducting standardised outcome analysis when conducting cost effectiveness analysis (Gold et al. 1996). There are various examples of instruments

that are utility based including the EQ-5D (Dolan 1997), the Health Utility Index (Torrance & Feeny 1989) and the Quality of Well-being Scale (Kaplan & Bush 1982).

Non-utility measures have also been called profile measures (Ribera et al. 2006), psychometric measures (Revicki & Kaplan 1993), non-preference based measures (Brazier et al. 2007) or health status measures (Tsevat et al. 1996). These measures involve deriving scores directly from the responses of subjects about their functioning and well-being. Responses from individual questions are then aggregated or summed into dimension-specific or cumulative scales with some known reliability and validity. It is possible for these measures to be either generic or specific (Revicki et al. 1995). Examples of such measures include the Barthel index; the medical outcomes study (MOS) scale and the Sickness Impact Profile (SIP).

When properly constructed, the psychometric measures have an attraction of having excellent reliability and validity. In addition, they are much easier to complete as there is minimal respondent burden, especially for shorter measures. Usually, a respondent is required to indicate the presence and frequency or intensity of symptoms, behaviours or feelings. The responses obtained are then aggregated into acceptable subscales or global scale scores (Revicki 1992).

8.3 Previous searches

A literature review of studies where a utility-based outcome measure was predicted from a non-utility-based measure was conducted in this chapter. Similar exercises were carried out by Revicki and Kaplan (1993), Brazier et al (2007) and Mortimer and

Segal (2008). Sections 8.3.1 to 8.3.3 present a summary of these studies. These three studies were identified as part of the search outlined in the sections 8.5 to 8.7.

8.3.1 Revicki and Kaplan (1993)

In this study, the primary focus was on establishing the evidence on the relationship between psychometric and utility-based approaches to the measurement of health related quality of life. A search was conducted in Medline. In total, 15 studies published between 1985 and 1993 were considered and these were studies where (1) HRQoL or health status scales were administered (2) health utility/ preference measures were used and (3) correlation or regression analysis was used to measure relationship between (1) and (2). Only four of the 15 studies (Revicki 1992; Tsevat et al. 1992⁶, Revicki et al. 1992 and Fryback et al. 1992) considered regression methods.

The key findings established by Revicki and Kaplan (1993) were that there is only moderate correlation between various health status measures and utilities. Regression analysis showed that the range for the R-squared statistic for predicting utilities from combinations of health status scales was between 0.18 and 0.43. The utility and non-utility measures were found not to be interchangeable indicators of health-related quality of life. In addition, psychometric health status scales were poorly to moderately correlated with Standard gamble and Time-Trade off utility scores.

8.3.2 Brazier et al (2007)

This was a systematic review of studies that reported the mapping or ‘cross-walking’ of generic preference based measures from non-preference based measures. Data were

⁶ This was an abstract which was developed into a paper later (Tsevat et al. 1996)

extracted from a total of 28 studies covering 119 different estimation models. A lot of non-preference and preference based measures were reported. In addition to searching fifteen databases, experts were also contacted regarding published and unpublished work.

The study found that the majority of studies were published after 2000 using data mostly from clinical trials. Higher R-squared or adjusted R-squared values (greater than 0.5) were reported for mappings of generic preference-based measures from generic non-preference based ones. Those involving the former and disease specific measures were more variable implying a higher degree of error in the models that considered these types of measures. Further work was needed in order to test the models reported in the reviewed studies in more relevant contexts and over a greater array of instruments.

8.3.3 Mortimer and Segal (2008)

This systematic review considered techniques for converting descriptive measures of health into quality-adjusted life year (QALY)-weights. Records from 1945 to 2006 were searched in a number of databases including citation databases, review article bibliographies and web-based resources. Four types of techniques were discovered: transfer to utility (TTU) regression, effect size translation, direct revaluation of an existing outcome measure using preference-based scaling methods and response mapping (RM). The group of techniques relevant to this chapter were those that used TTU regressions and RM - a total of 33 studies were reported. Most of these studies involved predicting utility measures from Short Form (SF)-family instruments.

In concordance with the results from Brazier et al (2007), the review found evidence that the TTU regression method was able to predict generic utility-based measures better from generic descriptive measures (with R-squared statistics of up to 0.71) than from disease-specific descriptive measures (R-squared statistics of up to 0.66). The RM techniques focussed on mapping individual to response levels on each item of a utility-based target measure. Most transformations predicting utility-based measures from non-utility based ones were ‘self-contained’ as they did not rely on additional data other than that from the measures in constructing the algorithms. Though useful conversion algorithms were found in the literature, there were a number of methodological issues that needed to be addressed in terms of the derivation and interpretation of predicted utility or QALY-weights.

8.4 Search Strategy

The same protocol employed in chapters two and three, based on Mugford (2001) and Roberts et al (2002), was utilised here. The search of databases carried out involved combining the terms shown in A6 in the appendix and searching for material from 1990 onwards. The following online computer databases were searched to identify relevant literature:

- ASSIA (CSA) – 1990 to 2008
- EMBASE (Ovid) 1990 – August 2008
- HMIC (Ovid) – 1990 to 2008
- MEDLINE (Ovid) 1990 – 2008
- ISI Web of Science, 1990 – August 2008
- PsycINFO 1990 – 2008

These searches encompassed literature relating to several utility and non-utility based outcome measures where explorations of relationships between such measures were carried out. The search was limited to literature published in English and was also not restricted to any country. A standard search string was devised comprising a combination of terms that described outcome measures. Publications were identified by their abstracts or, where these were unavailable, their titles and authors. Each publication was evaluated to determine its relevance and apparent importance to the review following the procedure shown in Roberts et al (2002). Where possible, copies were retrieved for more detailed evaluation. These articles, books or monographs were read and appraised and where possible, coded after the manner of Meads et al (2001).

8.5 How papers were selected for review

Stage I – Initial Scoping Search

A similar process as that conducted in chapter two was followed here. However, different criteria were used at this stage:

- Study design: Quantitative health and health related research.
- Population: any age.
- Setting: world-wide.
- Intervention: prediction of utility based outcome from a non-utility based one.

- Outcomes: All outcomes including health and functional outcome measures.
- Reporting: all studies excluding duplicates.

Stage II – Inclusion and exclusion criteria

A predetermined and explicit criteria was used to make the inclusion and exclusion decisions. Studies were only included if they were quantitative in nature. All studies that were qualitative in nature were excluded.

Stage III – Initial categorisation of studies

The process was again similar to that reported in chapter two except that the studies were subjected to the following criteria to determine their relevance to the literature review:

- A. Primary research is on direct prediction of one utility-based outcome measure using a predictor outcome measure that is not utility based.
- B. Emphasis on prediction of a utility based measure but the predictor outcome measure is utility-based as well.
- C. Primary research is on prediction of a non-utility measure from another non-utility measure or on their correlation.
- D. Contains useful information about outcome measures but there is no prediction involved or there no link to an already established utility based measure.

- E Study does not have any relevance to comparisons between outcome measures or predictions of outcome measures.

Using stage III criteria, studies that were coded as A were considered to be relevant to the review. Those classed as B, C, D or E were not deemed appropriate and were therefore not considered further.

Stage IV – Data extraction

Full papers were only read at this stage. All studies in category A were considered. The full papers or studies were read fully and information extracted into a table format. Full papers of some abstracts could not be obtained and these were not considered further. The exception was when the abstracts of such papers provided enough information needed in this review. The following information was recorded about each relevant study: primary focus of research, kind of population, sample size, study design, data sources, key outcomes and key findings (Table 8.1).

8.6 Results

Stages I, II, III, and IV

A total of 252 papers were initially selected from the initial scoping exercise (Stage I). At stage II, these studies were reduced to 145 papers, and to 54 at stage III upon reading the abstracts. During stage IV, full papers of most of the 54 studies were read and data was then extracted from 50 studies. This process is depicted in figure 8.1.

8.6.1 Description of studies

The search yielded a wide array of studies with publication dates ranging from 1992 to 2008. The population groups also differed in many aspects and these included those with stroke (van Exel et al. 2003), cancer (e.g. Ritvo et al. 2005, Chancellor et al. 1997 and Dobrez et al. 2007), general populations (e.g. Kulkarni, 2006, Nichol et al. 2001, Franks et al. 2004 and Gray et al. 2006), obesity (Brazier et al. 2004), dental problems (Brennan & Spencer 2006), Chronic obstructive pulmonary disease (Starkie et al. 2008) and knee pain (Barton et al. 2008a). Other conditions were arthritis (Bansback et al. 2007), intermittent claudication (Bartman et al. 1998), arterial disease (Bosch & Hunink 1996), diabetes (Barofsky et al. 2004), eye disease (Fryback et al. 1997), human immunodeficiency virus (HIV) related illness (Bult et al. 1998, Tsevat et al. 1996) and Asthma (Tsuchiya et al. 2002). Table 8.1 lists the other conditions. The sample sizes also varied with a minimum of 34 and a maximum of 37,933 observations. Of all the studies, 29 (about 57%) had less than 1000 subjects.

8.6.2 Utility measures evaluated

In total, ten utility-based measures were reported. The most commonly used was the EQ-5D utilised in 19 studies followed by the health utility index (HUI), in various versions, reported in 18 studies. Standard gamble (SG), visual analogue scale (VAS) and time trade-off (TTO) utilities were each used in 10 studies while the Short Form 6D (SF-6D) was utilised nine times. Also reported were the Quality of Wellbeing (QWB), used in eight studies, rating scale (RS) utilities used in four studies and the Assessment of Quality of Life -AQoL used in two studies (Mortimer et al. 2007 and Segal et al. 2004).

8.6.3 Non-utility measures evaluated

In contrast to the utility-based measures, there were 39 non-utility measures that were included in the studies reported in Table 8.1. While some of these were generic e.g. the SF-36 (Bartman et al. 1998), others were disease or condition specific e.g. Schizophrenia Quality of Life Scale – Revision 4 (SQLS-R4) reported in Clayson et al (2004), Inflammatory Bowel Disease Questionnaire – IBDQ (Buxton et al. 2007) and the Oral Health Impact Profile – OHIP-14 (Brennan & Spencer 2006). The most reported was the SF-36 found in 21 studies followed by the SF-12 which was used in nine studies. Nearly all of the other 37 measures were used once.

8.6.4 Types of analyses

All studies used, among other forms of analysis, regression to determine what variables were significant predictors of utility-based scores. All of the studies again used ordinary least squares (OLS) regression except Ritvo et al (2005) (Quintile regression), Bansback et al (2007) (generalised estimating equation), Brennan and Spencer (2006) (Tobit regression), Buxton et al (2007) (Maximum likelihood estimation) and Lawrence and Fleishman (2004) (best subsets multivariate linear regression). It is not clear what types of regression models were used by Yang et al (1997), Barton et al (2008a), Barofsky et al (2004), Chancellor et al (1997), Dixon et al (2003), Miller and Singer (2004), Segal et al (2004) and Tsevat et al (1996). In other studies, utilities predicted using already existing predictive algorithms, nearly all based on OLS regression models, were compared. These studies reported the results of all the regression algorithms with the choice of the best regression algorithm having to be made afterwards. These studies included Hollingworth et al (2002), Kaplan et al (2005), Lobo et al (2004), McDonough et al (2005), Pickard et al (2005),

Sherbourne et al (2001) and Thein et al (2005). Other studies did not use already existing algorithms but still compared the results of using different regression models to predict utilities. The regression models compared included OLS, Tobit and CLAD (Sullivan & Ghushchyan 2006a), OLS, ordered Probit and Interval (grouped data) (van Doorslaer & Jones 2003 and Lauridsen et al. 2004), OLS and GLM (Clayson et al. 2004) and OLS and multinomial logistic regression (Gray et al. 2006, Ouellet et al. 2008 and Tsuchiya et al. 2002). Other non-regression-based methods for predicting utilities have been used elsewhere (Mortimer & Segal 2008). Brazier et al (2002) however submit that econometric methods were more suitable in their work than techniques based on non-econometric methods such as multi-attribute utility theory (Torrance et al. 1982) because the dimensions of the SF-6D were not strictly independent.

As the focus of the studies was to predict utility scores, some criteria were needed to gauge the statistical difference and relationship between actual and predicted utility scores. The Pearson correlation coefficient was used to measure linear association in a number of studies (e.g. Revicki 1992; Shih et al. 2006; Kulkarni, 2006; Fryback et al. 1997 and Tsuchiya et al. 2002) as was the Pearson product moment correlation coefficient (Hollingworth et al. 2002). To compare near equivalence of actual and predicted scores, studies used the (root) mean squared error (MSE) (e.g. Brazier et al. 2004, Gray et al. 2006, Bansback et al. 2007 and Tsuchiya et al. 2002), mean prediction error (Sullivan & Ghushchyan 2006a) and 'mean absolute percentage' (Buxton et al. 2007). Also used were the mean absolute error (MAE) (e.g. Ouellet et al. 2008; Barton et al. 2008a), Forecast errors (Brennan & Spencer 2006), the Wilcoxon test (van Exel et al. 2003) and standard deviations (Lenert et al. 2005). To

compare performance of the models, some studies used the unadjusted or adjusted R-squared statistics (e.g. Bult et al. 1996 and Franks et al. 2003), C_p statistic (Lawrence & Fleishman 2004), Akaike Information Criterion – AIC (Clayson et al. 2004), inequality indices (Lauridsen et al. 2004) and the incremental cost utility ratio (Pickard et al. 2005). To test for differences in utility scores, the Kruskal-Wallis test (Bult et al. 1998) and Signed Rank tests (McDonough et al. 2005) were also used.

Some of the approaches employed in analysing relationships between outcome measures in a number of the 50 papers reported above have been adopted for the analysis that is reported in chapter nine.

8.6.5 Key findings

The significant message from the results of the 50 studies is that it is possible to predict utility scores from non-utility measures but this is not so in all instances. Despite this possibility, most of the studies show only poor to moderate correlation between utility and non-utility measures with very few studies showing high correlation. As a consequence, the R-square for the regressions varied from 0.14 (Bosch & Hunink 1996) to 0.85 (Bult et al. 1998). It also seems that utilities mapped from generic non-utility measures were better estimates than those mapped from disease-specific non-utility measures.

8.6.6 Methodological issues

There were many reasons given to explain why some regression models yielded low R-squares (i.e. the predictor outcome measure explained very little variation in the dependent variable outcome measure). The most significant reason advanced was that

these outcome measures were designed to measure different attributes (Tsuchiya et al. 2002). Revicki and Kaplan (1993, p.485) add that "...utility measures and psychometric health status measures are constructed to address different purposes." The former is designed to make a judgement about the impact of a health intervention on health outcomes while the latter seeks to discriminate among levels of functioning as well as to identify changes over time. Not accounting for heterogeneity in the modelling exercise (Bult et al. 1998) and focussing on individual rather than mean cohort utility values (Tsuchiya et al. 2002) were also causes of low R-squares.

Other methodological issues were identified in these studies. Many of the studies used OLS regression methods to predict utility scores from psychometric scores. The choice of this regression model however raises questions as it does not consider the ceiling effect of virtually all of the measures examined. Many of these results would therefore need to be checked against those obtained when a method that accounts for the ceiling effect, such as Powell's censored least absolute deviations (CLAD) estimator, generalised linear models (GLMs) or generalised estimating equations (GEE) are used. In the only study that compared the performance of the CLAD model to other regression models (Sullivan & Ghushchyan 2006a), fully specified CLAD model was found to be associated with the lowest mean prediction error.

Normality was assumed in some of the studies without any explicit mention of it being tested for (e.g. van Exel et al. 2004 and Revicki 1992). As a result, parametric rather than non-parametric tests were used in finding out the relationship between the utility and non-utility measures.

The choice of explanatory variables used in the regression analyses raises questions about what variables should be used in order to predict utility scores. Some studies used only the non-utility based measure as a sole independent variable (e.g. van Exel et al. 2004; Starkie et al. 2008; Bansback et al. 2007) while others like Tsuchiya et al (2002) included demographic variables in the list of explanatory variables. Clear guidance on this issue is essential.

Some studies such as van Exel et al (2004) only considered the summary or composite measures of the psychometric score in their model. But as Starkie et al (2008) and Tsuchiya et al (2002) have shown, unbundling the psychometric measures may unearth more relationships which would not be the case if a composite measure were used. Bartman et al (1998) went further to suggest the use of only those dimensions of the psychometric measure that were related to those on the utility measure.

The choice of the criteria for determining whether a regression model was a good fit in some cases causes concern. Some studies such as Nichol et al (2001) and Fryback et al (1997) used the R-square which has been shown elsewhere (Draper 1998) not to be the best statistic to use when testing for goodness-of-fit. It has been argued that the sum of squares due to error (SSE) and the root mean squared error (RMSE) statistics were better statistics for this purpose (Starkie et al. 2008; Tsuchiya et al. 2002; Bansback et al. 2007, Barton et al. 2008a). However, the cut-off points or levels for these statistics at which a judgement can be made on whether or not a model has a good fit is still debatable (Bansback et al. 2007).

Some sample sizes were too small for some of the regression results to be valid: Bartman et al (1998) had 34 subjects; Bosch et al (1996) had 68 while Revicki (1992) had 73. Care should therefore be taken when interpreting these results.

Lastly, most of the studies applied to a specific patient group and caution needs to be exercised when generalising results obtained from such samples to patients outside these patient groups (Barton et al. 2008a, Bansback et al. 2007). This is even more important when disease or condition-specific non-utility measures, which have been shown to be less good predictors of utilities, are used.

8.6.7 Predicting EQ-5D from Barthel scores

The literature review revealed only one study (van Exel et al. 2004) that predicted EQ-5D scores from Barthel index (BI). The study used a sample of 598 stroke patients in the Netherlands with a mean age of 73.5 years. The motivation for the exercise was to overcome the problem brought by missing values on EQ-5D which may obviously bias cost-effectiveness results.

Analysis of variance (ANOVA) was used to test whether mean EQ-5D scores differed among five Barthel categories defined as follows: independent (BI = 20), Mild (BI = 15-19), moderate (BI = 10-14), severe (BI = 5-9) and very severe (BI = 0-4). The study found that low EQ-5D scores were associated with low BI scores and this relationship was statistically significant. Ordinary least squares (OLS) regression method was used to predict EQ-5D scores from the BI at two and six months after stroke. Predicted EQ-5D scores were also obtained for the combined two-and-six-months-after-stroke data. Wilcoxon tests showed that predicted and observed EQ-5D

scores did not differ significantly from each other except for the ‘independent’ BI group. Intra class correlation (ICC) between observed and predicted EQ-5D values was however low for some BI categories. The study concluded that the following relationship exists between the EQ-5D and BI: the EQ-5D score is -0.25 for patients with a BI of 0, and increases by 0.05 for each additional point increase in the BI. Fully functionally independent individuals (BI = 20) were predicted to have an EQ-5D score of 0.75, which is in tandem with the general reference score for individuals in the age group considered (Kind et al. 1999). This study however did not also address the ceiling effects of both the BI and EQ-5D. It remains to be seen if the results would have been different if an alternative method had been employed. Also, the regression analysis only considered a single independent variable – BI. It would have been informative to investigate what the results would have been if the BI was unpacked or if other independent variables, especially demographic characteristics, such as age and gender were also considered. Lastly, and as conceded by the authors, the ICC was low to moderate as the predicted values were more clustered around the mean. Therefore, predicted values for individual patients can be misleading and the results are more useful only when considering cohorts of patients.

It will be therefore enlightening to see what kind of relationship would hold when the EQ-5D score is predicted from Barthel scores using the data from the national evaluation of costs and outcomes of intermediate care services for older people in the UK in chapter nine. Because of some methodological limitations identified in the studies in Table 8.1, the analysis in chapter nine addresses this question by:

1. Employing Powell’s censored least absolute deviations (CLAD).

2. Including demographic variables, specifically age and gender in the predictions.
3. Unbundling the BI outcome measure and using the various dimensions as potential predictors of EQ-5D.
4. Using stepwise regression to determine the variables that should be in the regression model.

8.7 Key messages for health services researchers from this chapter

The fundamental message from this chapter is that mapping from non-utility measures to utility scores using a regression framework is possible in some instances despite the poor to moderate correlation shown between the two types of measures in most studies. Also, the results appear to suggest that utilities mapped from generic non-utility measures were better estimates than those mapped from disease-specific non-utility measures. Only one study (van Exel et al. 2004) has predicted the EQ-5D from the Barthel index using a regression modelling framework based on a relatively small sample of individuals. The availability of a larger sample in the demonstration dataset therefore presented a unique opportunity for this kind of mapping exercise to be carried out in this thesis (chapter nine). Various regression models can be used for mapping exercises including OLS regression, Quintile regression, generalised estimating equation, Tobit regression, Maximum likelihood estimation and best subsets multivariate linear regression. Others are CLAD, ordered Probit and Interval (grouped data), GLM and multinomial logistic regression models. To measure goodness-of-fit, a number of statistics can be used including the R-square (adjusted and unadjusted), sum of squares error (SSE), mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), forecast errors, Akaike information criterion (AIC), various correlation coefficients and total prediction

errors. There is still considerable debate about the cut-off points or levels for these statistics at which a judgement can be made on whether or not a model has a good fit.

8.8 Conclusion

This literature review has revealed a number of regression approaches that have been used to predict utility measures from non-utility ones. Poor to moderate relationships were found between the two kinds of measures. The results of this literature review informed the empirical analyses that were conducted in the next chapter where the possibility of predicting the EQ-5D, a utility-based measure, from the Barthel index, a non-utility based measure, was examined.

Figure 8.1: Flow Chart of Literature Review

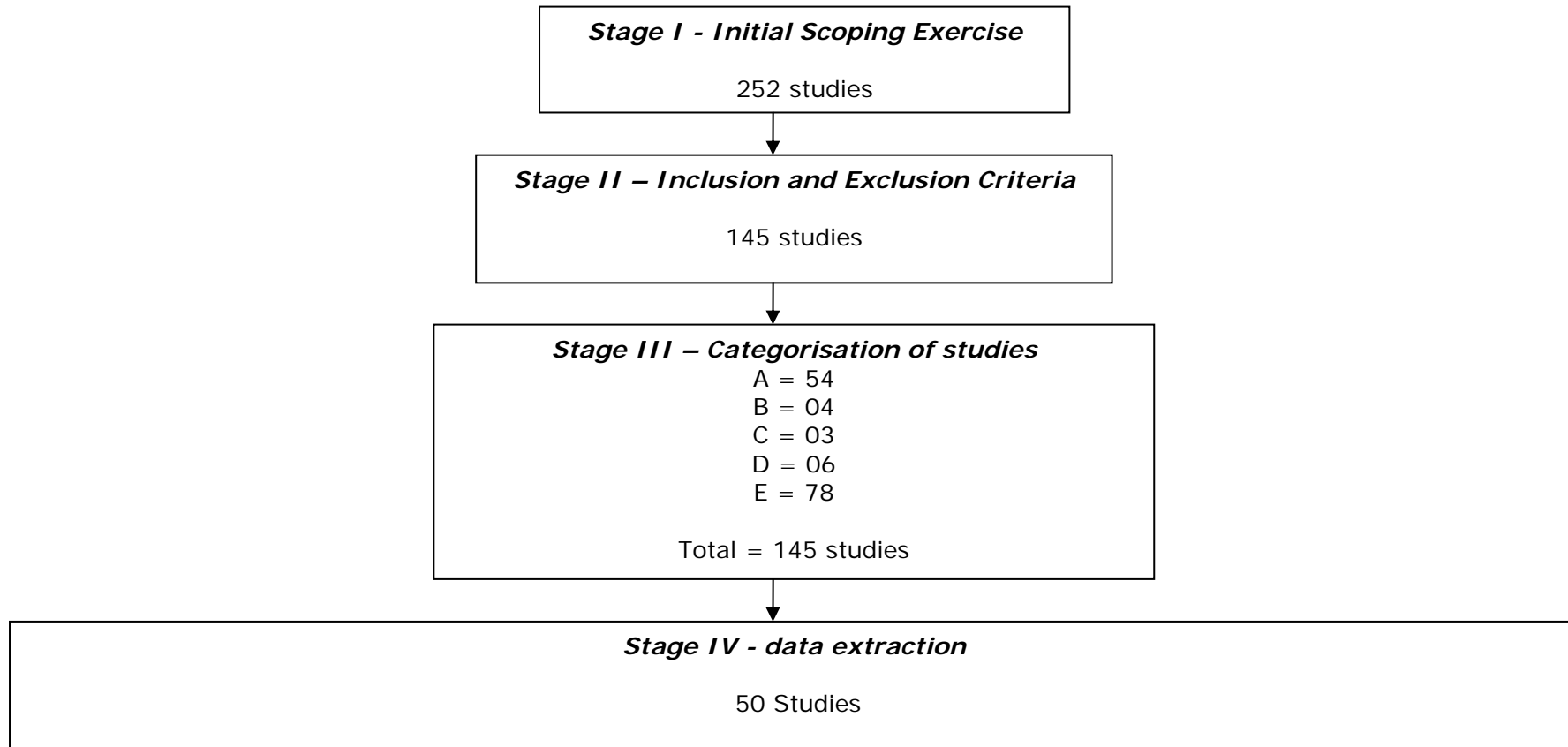


Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Bansback et al (2007)	A	Estimation of a preference-based single index (Eq-5D and SF-6D) using the health assessment questionnaire (HAQ)	470	Regression methods using generalised estimating equation algorithms, with the correlation matrix taking the structure of an autoregressive of order 1 [Root mean square error (RMSE)]	Patients with rheumatoid arthritis (RA) who were participating in 2 studies: one in the UK and another in Canada	HAQ	EQ-5D, SF-6D	The models were able to predict mean actual scores of cohorts for the EQ-5D and the SF-6D for the full range of the HAQ. The method may not however not the optimal one as certain aspects of the HAQ are not reflected in the EQ-5D and SF-6D.
Bartman et al (1998)	A	Among other objectives, the determination of which attributes of the medical outcomes study short form-36 (MOS SF-36) are significant predictors of health state valuation/preference (Health Utility Index – HUI)	34	Stepwise ordinary Least square (OLS) regression	Trial examining the efficacy of aggressive risk factor modification (ARFM) on functional independence of older people (>55) with intermittent claudication	MOS SF-36	HUI	Physical functioning and mental health best predicted HUI scores and R- square was 0.53. Only general health was a good predictor of the rating scale but the R-square was 0.59 No comparison between actual and predicted HUI2 scores was however conducted.
Bosch & Hunink (1996)	A	Assessment of the relationship between a descriptive measure (RAND 36-item Health Survey 1.0) and some valuations measures (Standard gamble – SG, Time trade-off – TTO, Rating scale – RS and the McMaster health utility index – HUI)	68	Stepwise ordinary least squares (OLS) regression with dimensions of the RAND-36 questionnaire as independent variables	Multicentre randomised clinical trial (Dutch Iliac Stent Trial) of the effectiveness of two percutaneous interventional procedures for patients with symptomatic peripheral arterial disease	RAND 36 questionnaire	SG, TTO, RS and HUI	Health status questions are not good predictors of SG utilities or TTO scores i.e. only explained 28 and 14% of the variation in the two measures, respectively. But they were better predictors of RS and HUI scores explaining 61 and 53% of the variation in these two, respectively

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Brazier et al (2004)	A	Estimation of a preference-based single index for the Impact of Weight on Quality of Life-Lite (IWQOL-Lite) instrument in obesity by mapping it onto the Short Form 6D (SF-6D) preference-based index.	1972	Stepwise ordinary least squares regression. [Adjusted R-squared and root of mean square error]	Community volunteers, participants in pharmacological and nonpharmacological clinical trials for obesity and gastric bypass surgery-bound individuals	IWQOL-Lite	SF-6D	Though certain aspects of the condition were not properly reflected in the SF-6D index, the mapping exercise produced useful results.
Brennan and Spencer (2006)	A	Mapping the 14 item Oral Health Impact Profile (OHIP-14) onto the EQ-5D.	375	Two Tobit regression models. [Forecast errors].	Random sample of South Australian dentists between 2001 and 2002	OHIP-14	EQ-5D	The continuous model performed better than the categorical one in terms of the mean forecast errors. It is possible for the OHIP-14 to be mapped onto the EQ-5D.
Bult et al (1996)	A	Examination of the relationship between standard gamble (SG) utility measure and the RAND-36 health status dimensions	68	Ordinary least squares (OLS); Latent class analysis. [R-squared]	Study involving 68 patients with symptomatic peripheral arterial disease	RAND-36	SG	When homogeneity was assumed, only 10% of the variation in SG could be explained by social functioning dimension. When heterogeneity was assumed, the within-R-squared was 0.12 for one class and 0.80 for another with an overall score of 0.49.

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Bult et al (1998)	A	Examination of the relationship between time trade-off (TTO) and the Short Form – 36 (SF-36)	263	Ordinary least squares (OLS) and Kruskal-Wallis one-way analysis of variance were used. Relationship examined assuming heterogeneity and then homogeneity, in turn.	Study involving 139 patients with various stages of human immunodeficiency virus (HIV) infection and 124 primary care control patients without HIV	SF-36	TTO	When homogeneity was assumed, only 33% of the variation in TTO could be explained by three dimensions (vitality, general health perception and social functioning). When heterogeneity was assumed, R-square increased to as high as 0.85 Assuming heterogeneity may improve explained variation in this kind of modelling
Buxton et al (2007)	A	Estimation of algorithms to map the Inflammatory Bowel Disease Questionnaire (IBDQ) and Crohn's Disease Activity Index (CDAI) onto Short Form 6D (SF-6D) and EQ-5D utilities.	3320	Regression models using maximum likelihood estimation. [R-squared and mean absolute percentage].	Clinical trials in Crohn's disease.	IBDQ and CDAI	SF-6D and EQ-5D	R-squared in IBDQ/EQ-5D model was 0.45 while that in the IBDQ/SF-6D model was 0.69. Algorithms are valid for estimating utilities from IBDQ scores but not from CDAI scores)
Chancellor et al (1997)	A	Derivation of EQ-5D and McMaster Health Utilities Index (Mark III) (HUI3) utilities from the Rotterdam Symptom Checklist (RSC) and the Hospital Anxiety Depression Scale (HADS).	98	Regression and factor Analysis	Clinical trial of two forms of radiotherapy for non-small cell cancer of the bronchus.	RSC and HADS	HUI3 and EQ-5D	It was not possible to use regression methods to map the RSC and HADS onto the HUI3 and EQ-5D

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Clayson et al (2004)	A	Predicting E-5D and Short Form 6D (SF-6D) from the Schizophrenia Quality of Life Scale- Revision 4 (SQLS-R4).	156	Ordinary least squares (OLS) regression and generalised linear model (GLM) with canonical log link. [R-squared, Akaike information criterion (AIC)].	Sample with Schizophrenia.	SQLS-R4	SF-6D and EQ-5D	The results from the three models were similar. EQ-5D OLS model with all items included had higher R-squared than the one with the domain scores.
Dixon et al (2003)	A	Derivation of utilities from Nottingham Health Profile (NHP) responses and Adult Growth Hormone Deficiency Assessment (AGHDA)	1433	Three regression models	Adults with growth hormone deficiency	NHP and AGHDA	Utilities	In general, the estimated models fitted the data well & discriminated between different patient populations and were also sensitive to changes in quality of life associated with treatment.
Dobrez et al (2007)	A	Estimation of an algorithm to convert responses to the Functional Assessment of Cancer Therapy - General (FACT-G) to mean time trade-off (TTO) utilities based on utilities for current health.	1433	Ordinary least squares regression with the constant constrained to one. [Mean prediction error].	Cancer patients randomly separated into construction and validation samples	Four FACT-G questions; Eastern Clinical Oncology Group - Performance Status (ECOG-PS) and Short-Form 36 (SF-36)	TTO	Algorithm predicted mean utilities well. The accuracy was comparable to that of other indirect preference-based measures of health-related quality of life.
Franks et al (2003)	A	Mapping the physical component summary (PCS-12) and mental component summary (MCS-12) of the Short Form 12 (SF-12) onto the EQ-5D and the Health Utilities Index (HUI3).	240	Ordinary least squares regression. [Adjusted R-squared].	Cross sectional data from a low-income, predominantly minority sample.	PCS-12 and MCS-12 of the SF-12	EQ-5D and HUI3	R-squared for EQ-5D model was 0.58 and 0.51 for the HUI3 model. Correlation between the two predicted utilities was high (0.96).
Franks et al (2004)	A	Mapping of physical component summary (PCS-12) and mental component summary (MCS-12) of the Short Form 12 (SF-12) onto the EQ-5D.	12,998	Ordinary least squares regression. [Unadjusted R-squared].	The 2000 US Medical Expenditure Panel Survey (MEPS)	PCS-12 and MCS-12 of the SF-12	EQ-5D	R-squared of about 0.63 were obtained for the both models. SF-12 component summary scales can be mapped onto preference scales though caution should be placed when dealing with lowest health states.

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Fryback et al (1997)	A	Using the Short Form 36 (SF-36) to predict Quality of Well-being (QWB) scores	1,356	A combination of stepwise and best-subsets Ordinary least squares regression employing polynomials in the eight Sf-36 profile scores to predict QWB scores. [Correlation between observed and predicted QWB values was calculated. The difference between the square of this correlation and the internal predictive model R-square was an indicator of the strength of prediction]	Data from the Beaver Dam Health Outcomes Study (BDHOS) which looked at eye-disease prevalence and risk factors. Adults over 45 years of age	SF-36	QWB	Up to 56.9% of the variation in the QWB was accounted for by a six-variable regression equation based on five SF-36 dimensions. SF-36 scores can therefore be used to predict QWB scores
Gray et al (2006)	A	Response mapping of Short Form 12 (SF-12) onto EQ-5D utility values.	12,967	Ordinary least squares (OLS) regression and Multinomial Logistic (ML) regression (Monte Carlo simulations). [Mean squared error - MSE and mean absolute error – MAE]	The 2000 US Medical Expenditure Panel Survey (MEPS)	SF-12	EQ-5D	ML models yielded higher MSEs but comparable MAEs to those of the OLS models.
Grootendorst et al (2007)	A	Determination of a model to estimate Health Utilities Index Mark 3 (HUI3) utility scores from Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) index scores	255	Ordinary least squares (OLS) and random effects regression methods. [Mean absolute error (MAE), root mean square error (RMSE) and the intra class coefficient (ICC)].	Data from a multicentre, open-label randomised clinical trial of patients with Osteoarthritis	WOMAC	HUI3	The model that had demographic variables as well as WOMAC pain, stiffness and function subscales was preferred with R-squared of 0.4. There were no statistically significant differences between the predicted and actual HUI3 utility scales at the group level.

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Hollingsworth et al (2002)	A	Comparison of utilities predicted from Short Form 36 (SF-36) responses using six different algorithms.	379	Regression methods based on Fryback et al (1997); Shmueli (1999); Brazier et al (1998); Lundberg et al (1999); Lawrence et al (1998) and Frybeck et al (1997); and Linert et al (2005). [Pearson's product moment correlation coefficient and effect size].	Cohort of low back pain patients who had been recruited to the Seattle lumbar imaging project (SLIP)	SF-36	Quality of well being (QWB); Standard gamble (SG); visual analogue scale (VAS) and time tradeoff (TTO).	Choice of SF-36 derived algorithm a determinant of the results. SG and TTO approaches generated higher preference values than the VAS.
Kaplan et al (2005)	A	Comparison of utilities predicted from Short Form 36 (SF-36) using four different methods.	636	Regression methods based on Fryback et al (1997); Nichol et al (2001) and Brazier et al (1998). [Pearson correlation coefficient, Fisher r to z transformation and analysis of variance].	Randomised clinical trial assessing a new treatment for adults with rheumatoid arthritis.	SF-36	Health Utilities Index (HUI2 and HUI3), EQ-5D and Health Assessment Questionnaire (HAQ)	Utilities from the four methods were significantly correlated to the measured HUI2, HUI3, EQ-5D and HAQ scores at baseline and at the end of the clinical trial
Koltowska-Haggstrom et al (2007)	A	Mapping the Quality of Life Assessment of Growth Hormone Deficiency in Adults questionnaire (QoL-AGHDA) onto the EQ-5D.	3,005	Multiple regression models	Swedish population	QoL-AGHDA	EQ-5D	The simple model without many demographic characteristics performed better than the full model.
Kulkarni (2006)	A	Conversion of numerical Hydrocephalus Outcome Questionnaire (HOQ) scores into health utility Index-2 (HUI-2) scores	140	Pearson correlation used to highlight association between HOQ and HUI-2 utility scores. The results were then used in an Ordinary least squares regression to transform HOQ scores into Utility scores	Responses of surgeons and mothers of children with Hydrocephalus attending the outpatient clinics at a paediatric hospital	HOQ	HUI-2	HOQ scores are readily translatable to HUI-2 utility scores using a simple transformation
Lauridsen et al (2004)	A	Mapping of Self-assessed health (SAH) onto the 15D		Comparison of 3 methods: Ordinary least squares (OLS) ordered Probit and interval (grouped data) regression models. [Inequality index].	Finnish Health Care Survey of 1995/1996.	SAH	15D	Interval regression model performs better than the other two.

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Lawrence and Fleishman (2004)	A	Predicting EQ-5D scores from the physical component summary (PCS-12) and mental component summary (MCS-12) of the Short Form 12 (SF-12)	14,580	2 variable regression model, best subsets multivariate linear regression. [C_p and R-squared statistics].	The 2000 US Medical Expenditure Panel Survey (MEPS)	SF-12	EQ-5D	2 variable model was the best with R-squared of 0.61. EQ-5D scores can be reasonably predicted from the Sf-12
Lenert et al (2005)	A	Comparisons of performance of a disease-specific utility mapping function for schizophrenia to a Short Form 36 (SF-36) -based utility mapping function.	474	Regression methods based on Positive and Negative Syndrome Scale (PANSS) and Nichol et al (2001) algorithms. [Standard deviation and differences between endpoint and baseline].	Data from a large, 1-year, open-label study of long-acting risperidone and to	Positive and Negative Syndrome Scale (PANSS) and SF-36	Health Utility Index Mark II (HUI2), Visual analogue scale and standard gamble.	Use of a disease-specific mapping function was practical. Variances in scaling and precision suggested that the disease-specific mapping function was better at detecting utility difference than the SF-36-based one.
Lobo et al (2004)	A	Comparison of utilities predicted from the SF-36 (Standard gamble SF-6D – SG-SF-6D, Quality of Wellbeing – QWB and Health Utilities Index Mark II – HUI2) to directly valued utilities.	99	Regression methods based on Fryback et al (1997), Nichol et al (2001) and Brazier et al (2002) methods. [Comparisons were based on Pearson correlations, Bland-Altman plots and analysis of variance].	Surviving lung or heart-lung transplant patients at the University of Minnesota Hospitals from November 1986 to January 1999.	SF-36	QWB, HUI-2 and SF-6D.	Findings supported the construct validity of all three methods. The Nichol and Brazier scores were generally higher than the Fryback scores.

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Longworth et al (2005)	A	Predicting EQ-5D scores from the Breathlessness Grade (BG), Canadian Cardio-vascular Society (CCS) patient demographic data.	510	Ordinary least squares regression. [R-squared].	Patients with stable angina in 4 UK cardiac out-patient clinics.	BG and CCS	EQ-5D	R-squared for final model was 0.37 and model predicted less severe angina better than severe angina.
Lundberg et al (1999)	A	Predicting rating-scale (RS) and time-tradeoff (TTO) utilities from the Short Form 12 (SF-12)	5,440	Ordinary least squares regression. [R-squared].	Postal questionnaire sent to 8,000 respondents in Uppsala County, Sweden.	SF-12	TTO and RS	R-squared in RS model was 0.482 and 0.232 in the TTO model. It is possible to convert the SF-12 into utilities.
McDonough et al (2005)	A	Comparison of utilities predicted from the SF-36 (Quality of Wellbeing – eQWB) to directly valuated utilities for EQ-5D and Health Utilities Index (HUI2 and HUI3).	2,097	Fryback et al (1997) regression method to obtain eQWB from the SF-36. [Spearman rank correlations and Signed Rank tests used to compare utilities].	Cross-sectional baseline data from an ongoing prospective study of interventions for back or leg symptoms associated with lumbar spine disorders (SPORT)	SF-36	EQ5D, HUI2, HUI3, eQWB, VAS, SF-6D	No instrument was superior. Measurement tool chosen should be one that best fits the condition under investigation as well as the study design.
Melsop et al (2003)	A	Translating the RAND Mental Health Inventory (MHI-5), self-rating health status (SRHS), Canadian Cardiovascular Society Classification (CCSC) and Duke Activity Status Index (DASI) into time tradeoff (TTO) utilities.	553	Ordinary least squares regression. [R-squared].	Patients with coronary artery disease and angina or ischemia	MHI-5, SRHS), CCSC and DASI.	TTO	R-squared ranged from 0.287 to 0.318.

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Miller and Singer (2004)	A	Predicting Utilities based on Visual Analogue Scale (VAS) and time tradeoff (TTO) from four domains of health related quality of life - HRQoL (Physical Pain/Discomfort, Independence, Social Relationships, Psychological Aspects).	782	Multiple linear regression	Cross-sectional supervised self-administered survey of Caucasian Americans and African Americans	The Four Domains	VAS and TTO.	The results suggested that HRQoL and utilities measure different things.
Mortimer et al (2007)	A	Derivation of item-based, subscale-based, and scale-based algorithms for mapping Short Form 36 (SF-36) data into Assessment of Quality of Life (AQoL) utility score.	455	Ordinary least squares and weighted least squares regression.	Stratified sample of persons aged more than 16 years and resident in Victoria, Australia.	SF-36	AQoL	A functional form with fewer restrictions does not necessarily lead to a lower magnitude of error when predicting between-group differences.
Nichol et al (2001)	A	Estimation of the Health Utility Index mark II (HUI2) from the SF-36	6921	Ordinary Least Squares (OLS) regression. [R squared]	Cross sectional data from a sample of patients that were insured by South California Kaiser Permanente	SF-36	HUI2	The models explained 50.5% of the variation in the HUI2. No comparison between actual and predicted HUI2 scores was however conducted.
Ouellet et al (2008)	A	Mapping of the Modified Rankin Scale (mRS) onto the EuroQol EQ-5D	1283	Ordinary least squares (OLS) regression (relationship between EQ-5D tariffs and mRS score), multinomial logistic regression (relationship between EQ-5D dimension responses and mRS scores), [Mean squared error, mean absolute error, quantile-quantile plots]	The Oxford Vascular Study on stroke or transient ischemic attack patients	mRS	EQ-5D	Including age and gender in the OLS and multinomial logistic equations predicted better EQ-5D tariffs.

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Paltiel et al (2001)	A	Estimation of time tradeoff (TTO) utilities predicted forced expiratory volume in 1 second (FEV ₁ , % predicted)	100	Ordinary least squares regression.	Cross-sectional study of adults (≥ 18 years of age) with asthma in the Lexington, Kentucky area.	FEV ₁ , % predicted	TTO	It was possible to predict TTO utilities from the FEV ₁ , % predicted.
Pickard et al (2005)	A	Comparison of utilities predicted from the SF-36 and SF-12 (Standard gamble SF-6D – SG-SF-6D, Quality of Wellbeing – QWB, EQ-5D and Health Utilities Index (HUI2) to directly valuated utilities.	220	Regression methods based on Fryback et al (1997); Shmueli (2004); Brazier et al (2002); Brazier and Roberts (2004); Nichol et al (2001); Franks et al (2003); Franks et al (2004); Lundberg et al (1999); Lawrence and Fleishman (2004). [Incremental cost utility ratio (ICUR) used to compare algorithms].	Clinical trial of adults with Asthma.	SF-36	SG-Sf-6D, HUI2 and QWB.	Higher ICURs were produced by methods based on Brazier algorithms.
Revicki (1992)	A	Examination of the relationship between health utility and psychometric health status measures	73	Among others, Pearson product-moment correlations to assess association and ordinary least squares regression to determine predictors of categorical and standard gamble utility scores	Randomised controlled trial (RCT) of Chronically ill liver patients	Parts of the sickness impact profile(SIP) and the medical Outcomes Study (MOS) health survey scales; Centre for Epidemiological Studies Depression (CESD) scale, Sexual dysfunction scale (SDS)	Torrance & Feeney's categorical rating and standard gamble (SG) utility measures	Only up to 27% of the variance in Utility measures can be accounted for by health status measures.

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Segal et al (2004)	A	Predicting utilities based on the Australian Assessment of Quality of life (AQoL), a visual analogue scale (VAS) for pain, and the Western Ontario and McMaster Arthritis Index (WOMAC) from the Short Form 36 (SF-36),	303	Multiple linear regressions. [Adjusted R-squared].	Patients with osteoarthritis in Australia.	SF-36	AQoL, Vas and WOMAC.	The models had good explanatory power with R-squared of at least 0.63.
Sengupta et al (2004)	A	Predicting health utilities index mark III (HUI3) and visual analogue scale (VAS) utilities from the Short Form 12 (SF-12),	6,923	Ordinary least squares regression. [Mean absolute deviation, mean squared deviation, Confidence intervals and R-squared]	Patients under Southern Californian Kaiser Permanente.	SF-12	HUI3 and VAS.	R-squared ranged between 0.35 and 0.55. The resulted supported the hypothesis that HUI3 and VAS utilities could be predicted from the SF-12.
Sherbourne et al (2001)	A	Comparing utilities based on the visual analogue scale (VAS), standard gamble (SG) and time tradeoff (TTO) and quality of well being (QWB) predicted from Short Form 12 (SF-12) and Short Form 36 (SF-36).	1,356	Ordinary least squares and regression methods based on Brazier et al (1998); Lawrence et al (1998) and Fryback et al (1997); and Lenert et al (2005).	Randomised controlled trial of patients with current depressive symptoms.	SF-12 and SF-36.	Partners in care (PIC) SG (PIC_SG), PIC-TTO, PIC-rating, SF-12UID, QWB, Brazier's VAS and Braziers SG.	Directly elicited utility values were lower than the predicted ones. Choice of utility-generating method has impact on cost-effectiveness results.
Shih et al (2006)	A	Association between the M.D. Anderson Symptom Inventory (MDASI) and utility	249	Three ordinary least squares (OLS) regression specifications to determine predictors of utility [Spearman's rank correlation to assess association between individual symptoms and utility]	Cancer patients from Tianjin Cancer Hospital in China	MDASI	Standard gamble utility score derived from the SF-36	Significant negative relationship between total symptom score and Utility (But very low R-squared)

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Shmueli (1998)	A	Predicting a visual analogue scale (VAS) utility from the Short Form 36 (SF-36).	1,956	Log linear regression. [R-squared].	Urban Jewish Israelis aged 45-75 years.	SF-12	VAS	The lowest R-squared was 0.403. More holistic scales of the Sf-12 (general health and vitality) are the main determinants of VAS scores
Shmueli (1999)	A	Mapping the Short-Form 36 (SF-36) and a five-category perception of general health (excellent, very good, good, fair and poor) onto a rating scale (RS) value (utility).	2,030	Unrestricted ordinary least squares (OLS) regression, anchored OLS and Box-Cox regression. [R-squared].	A sample of 45-75 year old people representing an Israeli Jewish urban population.	SF-36	RS values	The 8 scales of the SF-36 were linearly related to RS values with R-squared of 0.51.
Shmueli (2004)	A	Derivation of the relationship between visual analogue rating scale (VAS) valuation and the Short Form 36 (SF-36).	1,918	Unrestricted ordinary least squares (OLS) regression, anchored OLS and Box-Cox regression. [R-squared].	A sample of 45-75 year old people in an Israeli Jewish urban population.	SF-36	VAS	The results obtained were similar to those in Shmueli (1999)
Starkie et al (2008)		Mapping the St George's Respiratory Questionnaire (SGRQ) to Utility values using the EQ-5D	6112	Ordinary least squares (OLS) regression [R-square, Root Mean Squared Error]	A Chronic obstructive pulmonary disease (COPD) clinical trial called Towards a Revolution in COPD Health (TORCH) trial	SGRQ	EQ-5D	A model that had SGRQ item scores entered as categorical independent variables and descriptive independent variables was a better model at predicting EQ-5D. RMSE was 0.1708
Sullivan and Ghushchyan (2006a)	A	Predicting EQ-5D scores from the Short Form 12 (SF-12).	37,933	Ordinary least squares regression, Tobit regression and censored least absolute deviations (CLAD) regression. [Total prediction error, mean prediction error and 95% confidence interval].	The 2000 US Medical Expenditure Panel Survey (MEPS)	SF-12	EQ-5D	Gauged by mean prediction error, the CLAD model performed best followed by the OLS and Tobit. An algorithm to predict EQ-5D from SF-12 was provided.

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Thein et al (2005)	A	Estimation of utilities from the Short Form-36 (SF-36) for Chronic Hepatitis C	19 papers with samples varying from 38 to 1219	Regression analysis based on methods reported in three studies: Nichol et al (2001), Fryback et al (1997) and Shmueli (1998)	Systematic review of published studies where direct health-related quality of life assessment was done using the SF-36.	SF-36	Health utilities index Mark II (HUI2); Health utilities index Mark III (HUI3); Quality of well being (QWB); Visual analogue scales (VAS); Standard gamble (SG) and EuroQol EQ-5D.	Nichol's method, SF-36 utilities were comparable to direct patient elicited utilities. Fryback's utilities derived from QWB were also comparable to VAS patient-elicited utilities. Shmueli's utilities from VAS were higher than patient-elicited VAS utilities. SF-36 derived mean utilities were either lower or higher than expert estimates but comparable to patient estimates.
Tsevat et al (1996)	A	Assessment of health values of patients infected with human immunodeficiency virus (HIV)	139	Among others: multiple linear regression to determine relationship between each of the health value (preference) scores and all the health status measures	Human immunodeficiency virus (HIV) infected patients recruited from Healthcare Associates which is part of the Beth Israel Hospital	Short Form 36 (SF-36), mental health inventory (MHI) and Dyspnoea-fatigue index (DFI)	Quality of well-being (QWB) scale, rating scale (RS) and time trade-off (TTO)	Predictors of TTO were SF-36 physical functioning and vitality with R-square of 0.39 Those for RS were SF-36 general health perception subscale, MHI depression and DFI explaining 52% of the variance QWB had SF-36 vitality and bodily pain, DFI and higher education (a demographic variable) as predictors. R-square was 0.71
Tsuchiya et al (2002)	A	Converting the Asthma quality of life questionnaire (AQLQ) into EQ-5D indices	3000	Ordinary least squares (OLS) and multinomial logistic regression analyses on 6+2 models. [Mean squared error – MSE and Pearson's correlation coefficients].	A randomised controlled trial on the effectiveness of computerised decision support in primary care for patients with Asthma	AQLQ	EQ-5D	The model with EQ-5D as a continuous dependent variable and AQLQ item levels as categorical independent variables is the best. It is possible to map AQLQ scores onto EQ-5D scores but this has to be for a cohort and not individual patients because of the large error associated with prediction of the latter. Mapping however should be seen as a 'poor second best' option

Table 8.1: Studies selected after Stage III

Study	Final Class	Primary Focus	Sample size	Type of Analysis [method used to gauge performance]	Data Sources	Key Outcomes		Key Findings
						Psychometric measures	Utility measures	
Van Doorslaer and Jones (2003)	A	Mapping ordinal responses on a self assessed health (SAH) question onto the Health Utilities Index mark III (HUI3).	15,539	Comparison of three methods: Ordinary least squares (OLS) ordered Probit and interval (grouped data) regression models. [Comparisons with benchmark HUI scores].	Canadian National Population Health Survey	SAH	HUI3	Interval regression model performs better than the other two.
van Exel et al (2003)	A	Prediction of the EQ-5D from the Barthel index (BI)	598	Ordinary least squares (OLS) regression. [Wilcoxon test, R-squared and spearman's correlation].	Data collected from the Evaluation of Dutch Integrated Stroke Service Experiments (EDISSE) study on elderly people 2 and 6 months after suffering a stroke	Barthel index	EQ-5D	The BI can be used as a proxy for missing EQ-5D scores. There is low to moderate correlation between the observed and predicted EQ-5D scores. The observed and predicted values also do not differ for practically all the subgroups
Yang et al (2007)	A	Mapping medical outcomes study (MOS) sleep scale onto a preference-based health state utility index Short Form -6D (SF-6D)	1735	Various ordinary regression models to assess the predictors of SF-6D	Three datasets (1) 4-year observational study of chronically ill patients (2) 7-week non-comparative clinical trial of an osmotic controlled-release oral system (OROS) hydromorphone in the treatment of chronic low back pain (3) 6-week open-label RCT of OROS hydromorphone in the treatment of pain associated with chronic osteoarthritis	MOS – Sleep Problem Index (SLP9)	SF-6D	The best fitting model explained 34% of variation in SF-6D Mean predicted and observed SF-6D scores were nearly identical It is possible to map MOS SLP9 onto SF-6D scores

CHAPTER NINE - CHALLENGES WITH OUTCOME DATA: EMPIRICAL ANALYSIS

9.1 Introduction

In the national evaluation of costs and outcomes of intermediate care services for older people in the UK, data on the EQ-5D and the Barthel index were collected. These two measures were described in detail in chapter four. This chapter addresses the question of whether missing values on the EQ-5D can be predicted from the Barthel index. In van Exel et al (2004), missing values of the EQ-5D were predicted from the overall Barthel index score which was used as the sole independent variable in a regression model framework. Other studies, as shown in chapter eight, have also used regression prediction approaches to predict preference or utility-based outcome measures from non-utility-based ones. This chapter adopts some of these approaches and reports the results of an empirical analysis which examined the possibility of predicting the EQ-5D, a utility-based measure, from the Barthel index, a non-utility-based conventional clinical scale of functional independence measure. Therefore, the focus of the analyses has not been on explaining the relationship between the two types of measures but rather on the prediction of one from another. Key messages to health services researchers who may be working with these or similar outcome measures are also presented in the discussion.

9.2 Methods

9.2.1 Data

Data were obtained from the national evaluation of costs and outcomes of intermediate care services for older people in the UK described in chapter four. Data on patient characteristics, descriptors of the intermediate care services, and

intermediate care-related services' descriptors were collected. These data included the two health outcome measures - the EQ-5D and the Barthel which were described in detail in section 4.5 of chapter four. Data on the outcome measures were collected at admission and discharge.

9.2.2 Statistical Analyses

The statistical analyses in this chapter were in three different stages.

In the first stage, descriptive summary statistics were generated for all the main variables. Demographic and baseline characteristics of the sample were generated. A visual inspection of the relationship between the total EQ-5D and Barthel index scores was done using scatter diagrams. The relationship between the total EQ-5D score and five 'categories' of the Barthel was examined using a box-plot. The Barthel index categories were adapted from the Wade et al (1988) classification. Five Barthel index categories were considered: independent (Barthel index = 20); mild (Barthel index = 15-19); moderate (Barthel index = 10-14); severe (Barthel index = 5-9); and very severe (Barthel index = 0-4). Box plots showing the relationship between the EQ-5D dimensions and the total Barthel scores were also produced.

In the second stage, the model arrived at by van Exel et al (2004) and described in chapter eight was tested on the demonstration dataset to see how well it predicted EQ-5D scores. This model was used because it came from the only study known to date to have predicted EQ-5D tariff scores from Barthel index scores. The model was depicted algebraically as:

$$EQ-5D = -0.25 + 0.05 \text{ Barthel} \quad (9.1)$$

From equation (9.1), we can deduce that a totally dependent individual (Barthel = 0) has a corresponding EQ-5D score of -0.25, which is worse than death while a Barthel score of 5.0 is equivalent to death (EQ-5D = 0). A totally independent individual (Barthel = 20) has a corresponding score of 0.75 on the EQ-5D. This is comparable to the reference value of the general population in the UK for ages 65-80 (Kind et al. 1999).

In the last stage, eleven different regression model specifications were analysed to see if they performed better than the van Exel et al (2004) model. The overall EQ-5D tariff score or dimensions of the EQ-5D were the dependent variables and the overall Barthel score or Barthel dimensions were entered as independent variables. In some models, demographic characteristics of patients were included as independent variables. The appropriate model specification was not clear a priori hence a number of the specifications were employed underlying the exploratory nature of this analysis. This approach has been used elsewhere (Brazier et al. 2002 and Busschbach et al. 1999). There were some assumptions that needed to be made about these regression models but it was not necessary for all of them to hold for each of the models:

- a) that the Barthel items within a given dimension carry equal weight.
- b) that the Barthel and the EQ-5D both cover all the domains of health of relevance to the population in consideration.

- c) that the 10 Barthel dimensions carry equal weight.

The eleven model specifications were⁷:

- (1) Overall EQ-5D tariff score = $f(\text{Overall Barthel score, age, gender})$.

The overall Barthel index, age and gender were independent variables. The EQ-5D, Barthel score and age were entered as continuous variables while gender was entered as a dummy variable. This model requires all three assumptions.

- (2) Overall EQ-5D tariff score = $f(\text{Overall Barthel score})$.

This was a reduced version of model (1) where the overall Barthel score was the only independent variable. This model requires all three assumptions.

- (3) Overall EQ-5D tariff score = $f(\text{Barthel dimensions, age, gender})$.

To obtain this model, a regression where all the 10 dimensions of the Barthel index including age and gender were included as independent variables was run. All variables except gender were treated as continuous variables. This requires assumptions (a) and (b).

- (4) Overall EQ-5D tariff score = $f(\text{Barthel dimensions} - \text{continuous variables})$.

Age and gender were excluded as covariates from model (4) but all 10 dimensions of the Barthel index were included in this regression model. All variables were treated as continuous variables. This requires assumptions (a) and (b).

⁷ Algebraic representations of these models are as given in section 5.4.1 of chapter five. As before, 'Overall EQ-5D tariff score = $f(X)$ ' indicates that the overall EQ-5D tariff score is some function of X .

(5) Overall EQ-5D tariff score = $f(4 \text{ Barthel dimensions} - \text{continuous variables})$.

In model (3), only four dimensions were significantly related with the EQ-5D. Age and gender were not significant. In model (5), a model was specified with only four dimensions of the Barthel Index used as continuous independent variables. These were “Transfer”, “Mobility”, “Dress” and “Stairs” chosen after running a stepwise regression on all 10 dimensions of the index plus age and gender. This model needs assumptions (a) and (b).

(6) Overall EQ-5D tariff score = $f(\text{Barthel dimensions} - \text{continuous variables}, \text{Interaction terms})$.

This model was similar to the one immediately before with the exception that the independent variables also included interaction terms. These interaction terms were identified from model (4) and selected if their coefficient was positive (showing that there was a positive relationship between EQ-5D and Barthel, as expected) and the coefficient was significant at 5% level. This model requires assumptions (a) and (b).

(7) Overall EQ-5D tariff score = $f(\text{Barthel dimensions} - \text{categorical variables}, \text{age, gender})$.

Here, all the 10 dimensions of the Barthel index including age and gender were entered as independent variables of a regression model. While the overall EQ-5D tariff score was considered to be continuous, all the independent variables with the exception of age were entered into the equation as categorical variables. For the Barthel dimensions, the ‘dependent’ level was used as the reference category. Assumption (b) was presumed for this model.

(8) Overall EQ-5D tariff score = $f(\text{Barthel dimensions} - \text{categorical variables})$.

Model (8) was a reduced version of model (7) in which the demographic variables (age and gender) were removed. All the 10 dimensions of the Barthel index were treated as categorical variables with the 'dependent' level again used as the reference category in the regression analysis. Assumption (b) was presumed for this model.

(9) Overall EQ-5D tariff score = $f(4 \text{ Barthel dimensions} - \text{categorical variables})$.

The same reasoning that was used to define model (5) was followed here as well. Model (9) was a reduced version of model (8) and was arrived at using stepwise regression. Here again, all the dimensions of the Barthel index were entered as categorical variables while the overall EQ-5D tariff score was considered to be continuous. A regression model was used again and the 'dependent' level of each of the Barthel dimensions was treated as the reference category in the analysis. Assumption (b) was presumed for this model.

(10) EQ-5D dimensions = $f(\text{Barthel dimensions} - \text{categorical variables})$

The EQ-5D was entered as a categorical variable. As a result, five different sets of regressions were run i.e. one for each of the five EQ-5D dimensions. Category one for each of the EQ-5D dimensions (no problems) was used as the base category. The independent variables were treated as categorical variables. The 'dependent' level of each of the Barthel dimensions was used as the reference category in the analysis. This model requires assumption (c).

(11) EQ-5D dimensions = f (Overall Barthel score)

Model (11) was another regression model which also resulted in five different sets of regressions being run. The 'No Problems' category for each of the each of the five EQ-5D dimensions was again used as the base category in the analysis. Only one independent variable, the overall Barthel score was used as an independent variable and it was treated as a continuous variable. This model requires assumption (c).

9.2.3 Choosing a regression model family

A number of regression model families were considered for use in the models (1) to (9). These were ordinary least squares (OLS), generalised linear models (GLMs), random or fixed effects models, Tobit models and Powell's Censored Least Absolute Deviations (CLAD) estimator.

The overall EQ-5D tariff score which was used as the dependent variable has an upper bound of 1.0. OLS regression models have provided an easy method for predicting utility values from non-utility-based instruments in a number of studies (e.g. van Exel et al. 2003; Tsuchiya et al. 2002). But the use of such a method has the potential limitation in the presence of ceiling effects which has been shown to produce inconsistent estimates of the coefficients of independent variables (Gray et al. 2006; Long 1997). In the national evaluation of intermediate care dataset, 7% of the EQ-5D scores reached the ceiling score of 1.00. When the method was tried on the demonstration dataset, predicted EQ-5D tariff values of more than 1.00 were obtained for some observations. Truncating all values above 1.00 to 1.00 has however been used by others and so this method was chosen as one of those to be used.

An alternative was to use a GLM model in which the dependent variable is transformed into an s-shaped non-linear variable (McCullagh & Nelder 1989). This transformation ensures that the dependent variable approaches the value of 1.00, but does not reach it. In the demonstration dataset, however, there are many observations whose value for the EQ-5D was 1.00 and this method would therefore be biased against these observations as they would be dropped from the analysis. Tsuchiya et al (2002) proposed standardising the raw EQ5D indices to the range [0, 1], based on an artificial range and transforming this later. This would however be not the best way to proceed as even this crude transformation would lead to crude estimates.

Though the EQ-5D and Barthel scores were collected at two points (at admission and discharge), no repeated observations were used in the analysis. Therefore there was no theoretical basis to use either a random or fixed effect models.

The Tobit model has been shown to be a viable alternative for analysing data with a ceiling effect (Greene 1997). However, biased estimates would be obtained in the face of heteroscedasticity or nonnormality. The Breusch-Pagan test ($\chi^2 = 60$, $p < 0.001$) and the Shappiro-Francia test (z-score = 15.1, $P < 0.001$) confirmed the presence of heteroscedastic and skewness, respectively. It was therefore not appropriate to use the Tobit model.

The CLAD regression model takes the ceiling effect into account and calculates bootstrapped estimates of the sampling variances as opposed to analytical standard errors (Chay & Powell 2001; Clarke et al. 2002). Added advantages of the CLAD estimator are that it is robust to heteroscedasticity, it is also consistent and

asymptotically normal for a wide class of error distributions (Arabmazar & Schmidt 1981; Vijverberg 1987) and can be used to model data with skewed distributions (Johnston 1997; Huang et al. 2008). It estimates the regression coefficients so as to minimize the sum of the absolute value of deviations from the regression line (Austin 2002). The CLAD has also been shown by Sullivan and Ghushchyan (2006a) and Austin (2002) to produce lower prediction errors than the OLS.

As a result, the decision was made to use CLAD and OLS model for models (1) to (9). The OLS model has been described in greater detail in section 5.5.4.

The CLAD model is similar in specification to the Tobit model and can be depicted as follows (Greene 1997):

$$EQ-5D_i^* = f(X_i) + \mu_i \quad (9.2)$$

where:

$EQ-5D_i^*$ = individual i 's true EQ-5D score,

X_i = vector of individual i 's characteristics and

μ_i = stochastic error term for the i^{th} individual.

Then $EQ-5D_i = 1.00$ if $EQ-5D_i^* \geq 1.00$ and $EQ-5D_i = EQ-5D_i^*$ if $EQ-5D_i^* < 1.00$.

The CLAD model describes the association between the median EQ-5D and independent variables (Powell, 1984; Huang et al. 2008). Therefore, the CLAD model predicts median EQ-5D utility scores conditional on the predictors for each subject.

The relationship was chosen because the median is considered to be more robust than the mean to ceiling effects and skewness (Powell, 1984; Chay & Powell 2001; Clarke et al. 2002; Johnston 1997). The mean presented is therefore mean of the predicted conditional medians for the subjects. This approach has been used before in other studies (Sullivan & Ghushchyan, 2006a; Cheung et al. 2008; Payakachat et al. 2009; Cheung et al. 2009).

For models (10) and (11), there was the option of using the ordinal logistic regression or multinomial logistic regression since the dependent variables were treated as categorical variables with discrete outcomes. The goal was to predict the probability of each response level for each of the five EQ-5D dimensions. When using ordinal logistic regression, the order of the responses to each of the EQ-5D questions/dimensions is taken into account. The model also derives the relationship between an unobserved or latent continuous variable and the thresholds or cut points of the observed grouped variable (Gray et al. 2006). The assumption of proportional odds or parallel regression however needs to be met for the results of the model to be accurate. Running the ordered logistic model and testing using the Wald test revealed that this assumption was violated.

Multinomial logistic regression models were therefore used for models (10) and (11). These models can be depicted in the following way (Greene 1997; Long 1997):

$$\Pr(EQ-5D_Dimension_i = m / X_i) = \frac{\exp^{X_i \beta_m}}{\sum_{j=1}^J \exp(X_i \beta_j)} \quad (9.3)$$

where:

$\Pr(EQ-5D_Dimension_i = m / X_i)$ = the probability of observing outcome m for a particular EQ-5D dimension given X for individual i and

$X_i\beta_j$ = a linear combination of individual i 's characteristics (X_i) and logit coefficients (β_j).

9.2.4 Robustness

Two approaches were used to measure the predictive accuracy of models. As pointed out earlier, the EQ-5D and Barthel scores were collected at two points: at admission and at discharge. In the first approach, therefore, the admission dataset was used as the 'prediction dataset' while the discharge dataset was treated as the 'validation dataset'. This choice was made purely on the basis of the sample size: the admission dataset had about 31% more observations than the discharge dataset. The prediction dataset was made up of 1189 observations while 910 observations were used in the validation dataset. The prediction dataset was used for running all the models. Models estimated from the prediction dataset were then used to predict EQ-5D tariff scores from the validation dataset ("out-of-sample" validation). The validation dataset was thus used to predict EQ-5D tariff scores using the regression coefficients from the prediction dataset. This allowed for the models to be tested on their ability to predict utility scores of "out-of-sample" subjects. "Out-of-sample" subjects are those whose data were not used to estimate the models (Kennedy & Aoki 2002). However, this approach has the problem that the validation dataset is not strictly independent of the prediction dataset as it includes many of the patients who are also in the prediction dataset.

For this reason, a second approach was also considered. This involved randomly splitting the admission dataset into two so as to have another ‘prediction dataset’ and another ‘validation dataset’. The prediction dataset was used in the same way as above i.e. to produce ‘within-sample’ predictions while the validation dataset was used to obtain the ‘out-of-sample’ predictions. Predictions obtained from the validation dataset were used for “out-of-sample” validation. The prediction dataset was made up of two-thirds of the admission dataset (793 observations). The remaining third of the complete dataset (396 observations) was used as the validation dataset. A similar split has been used in analogous exercises (e.g. Bansback et al. 2007).

To measure goodness of fit, the root mean squared error (RMSE) and the mean absolute error (MAE), which can all be viewed as measures of the difference between the predicted and observed values of the dependent variable (EQ-5D), were used. The RMSE is a measure of predictive error - the difference between the predicted value and the actual observed value (Daniel & Terrell 1995). It is the square root of the mean squared error (MSE) and is often preferred to the MSE because it is on the same scale as the data (Hyndman & Koehler 2006). A lower RMSE for a predictive model implies that it produces more accurate predictions and is therefore a better performing predictive model. To obtain the RMSE, the difference between the observed and predicted values of the EQ-5D tariff scores was squared and then summed over all observations. The mean of these summed values is the MSE, which was square-rooted to obtain the RMSE. The MAE is another measure of predictive error. It is the mean of the absolute values of the difference between the predicted value and the actual observed value (Daniel & Terrell 1995). The MAE was calculated by summing the absolute difference between the observed and predicted values of the EQ-5D tariff

scores over all observations. The mean of the summed values is the MAE. A smaller MAE implies a better predictive model. Algebraically, these two measures of error can be represented as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (9.4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (9.5)$$

where:

n = the number of observations (number of individuals in the sample),

y_i = the value of the observed EQ-5D score for the i^{th} individual,

\hat{y}_i = the value of the predicted EQ-5D score for the i^{th} individual.

The RMSE is however more sensitive to outliers than the MAE (Hyndman & Koehler 2006).

In addition, the ranges of the predicted and observed EQ-5D scores as well as the Spearman correlation (Altman 1991) between them were calculated.

9.2.5 Predicted EQ-5D indices from models (10) and (11)

The predictions generated from the multinomial logistic regressions (models 10 and 11) represent the probability that a particular individual has a given level on each of the EQ-5D dimensions. For instance, these regression models can predict the

probability of individual 'A' having level 1, 2 or 3 on the 'Mobility' dimension. There are three methods of obtaining predicted overall EQ-5D tariff scores from the results of a multinomial logistic regression.

Tsuchiya et al (2001) refer to the first one as an "indirect" way. Here the level for each of the five EQ-5D dimensions associated with the highest probability is chosen as representing the level for an individual. Once the levels for each of the five dimensions are determined, the EQ-5D tariff score can be determined from the EQ-5D valuations. However, this method does not utilise all the information as it does not take into account the magnitude of the differences in the probability i.e. a probability split of 44%-43%-13% for three levels of a given dimension would be treated the same as one of 92%-5%-3%. Gray et al (2006) also assert that this method would result in biased expected EQ-5D tariff values being obtained.

Another method referred to by Tsuchiya et al (2001) as the "direct" approach combines the probabilities for different dimensions and levels with the population value set to calculate an expected EQ-5D index. This method however has the disadvantage of not being able to obtain a predicted overall EQ-5D tariff score of exactly 1.00 most times. As there are a number of individuals in the demonstration dataset who had an EQ-5D value of 1, the second method was not considered to be appropriate.

Gray et al (2006) suggest a third method of using Monte Carlo simulation to estimate a predicted response category for each item of the EQ-5D for each individual in the sample. Random numbers are drawn and compared to the predicted probability from

the multinomial logit that an individual will be in a particular health state. It is then possible to fit individuals into EQ-5D descriptive states which situation would enable one to apply the EQ-5D valuations to obtain an EQ-5D tariff score. This method reduces the bias that can be obtained from the first method and also does not suffer from the same pitfalls as those of the second method. Box 9.1 below illustrates the Monte Carlo procedure. In a similar exercise involving 12,967 observations, Gray et al (2006) carried out only one simulation and found this to be adequate. They however point out that more simulations are needed for smaller samples. Since the sample used for models (10) and (11) was about one-tenth of the Gray et al (2006) sample, eleven multiple simulations were run in order to get the predicted EQ-5D scores.

Both the first and third methods were used in the prediction of EQ-5D tariff scores from multinomial logistic regression models and their results were compared. STATA version 8.2 (StataCorp LP, 2004) was used for all analyses.

Box 9.1: Monte Carlo Method for obtaining predicted EQ-5D indices from multinomial logistic regression models

Consider individual 'A' for whom the predicted probabilities from a multinomial logistic regression output on the Mobility dimension of the EQ-5D are p_1 (No problems), p_2 (Some problems) and p_3 (Extreme problems), then the following will be the procedure:

1. Random variable X is generated from a uniform distribution i.e. $X \sim U[0, 1]$.
2. If :
 - $X < p_1$, then select 'No Problems',
 - $p_1 < X < (p_1 + p_2)$, then select 'Some Problems,
 - $(p_1 + p_2) < X < 1$, then choose 'Extreme Problems'.
3. Repeat steps (1) and (2) for each of the other four EQ-5D dimensions.
4. Obtain an EQ-5D profile (descriptive states) for individual A and then apply the EQ-5D evaluations to get an EQ-5D tariff score.
5. Repeat steps (1) to (4) eleven (11) times, using a different seed each time for drawing up the random number from the uniform distribution.
6. Calculate the average EQ-5D for individual 'A' by summing up all 11 EQ-5D scores and dividing by 11.

9.3 Results

9.3.1 Demographic characteristics

A sample of 1189 intermediate care clients was included in the analysis (Table 9.1).

The mean age was 79.4 and the majority (70%) were female. At the time of inception

into an intermediate care service, 54% were living alone. The mean EQ-5D and Barthel scores at admission were 0.42 and 14.8, respectively. The corresponding mean scores at discharge for these two measures were 0.61 and 16.7. Most intermediate care schemes were performing either an acute admission (55%) or supported discharge (42%) function.

Table 9.1: Sample demographic and baseline characteristics

n	1189
Age - Mean (SD)	79.4 (10.34)
% Male	30
% Living Alone	54
EQ-5D at Admission - Mean (SD)	0.42 (0.36)
Barthel at Admission - Mean (SD)	14.8 (4.22)
% Admission Avoidance	55
% Supported Discharge	42
% Residential Intermediate Care	18

9.3.2 Relationship between EQ-5D and Barthel scores – Raw data

Figure 9.1 below depicts the relationship between the overall EQ-5D tariff scores and the Barthel overall score at admission. No clear pattern is discernable from the scatter plot. One reason for this is the fact that some of the dots represent more than one observation. If however the Barthel index is split into categories of ‘very severe’ (Barthel = 0-4), ‘severe’ (Barthel = 5-9), ‘moderate’ (Barthel = 10-14), ‘mild’ (Barthel = 15-19) and ‘independent’ (Barthel = 20) functioning as shown in figure 9.2, then a positive relationship is evident in that lower (higher) Barthel scores are associated with lower (higher) EQ-5D scores.

Figure 9.1: Scatter plot of composite EQ-5D and Barthel scores at admission

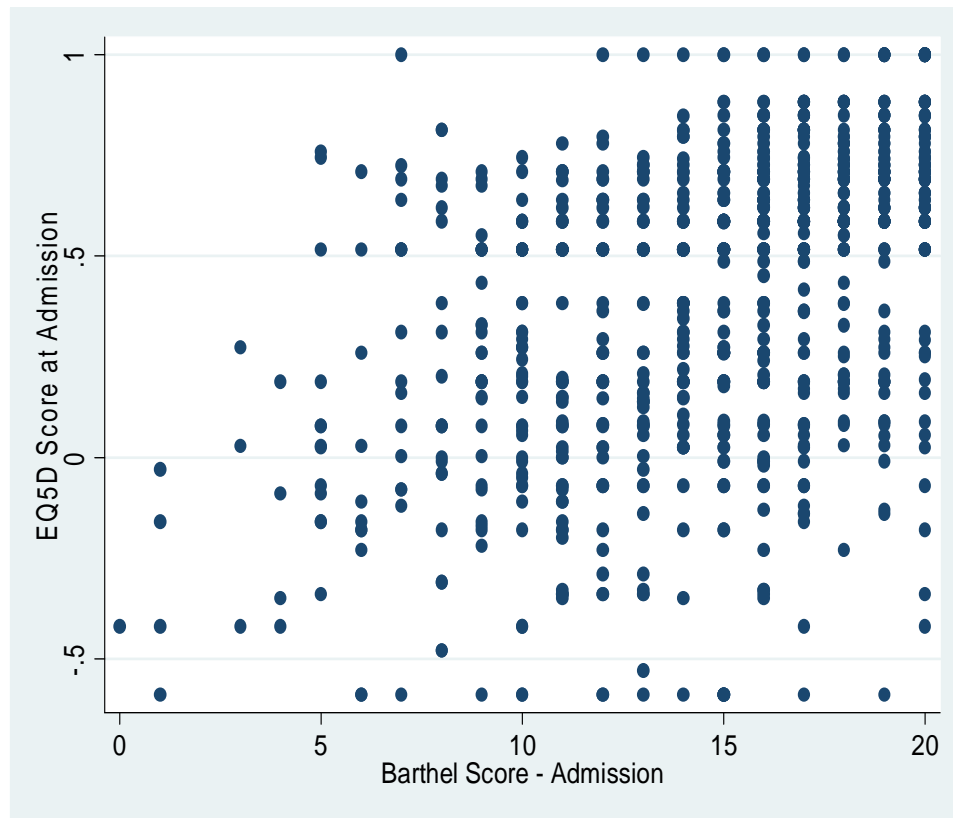
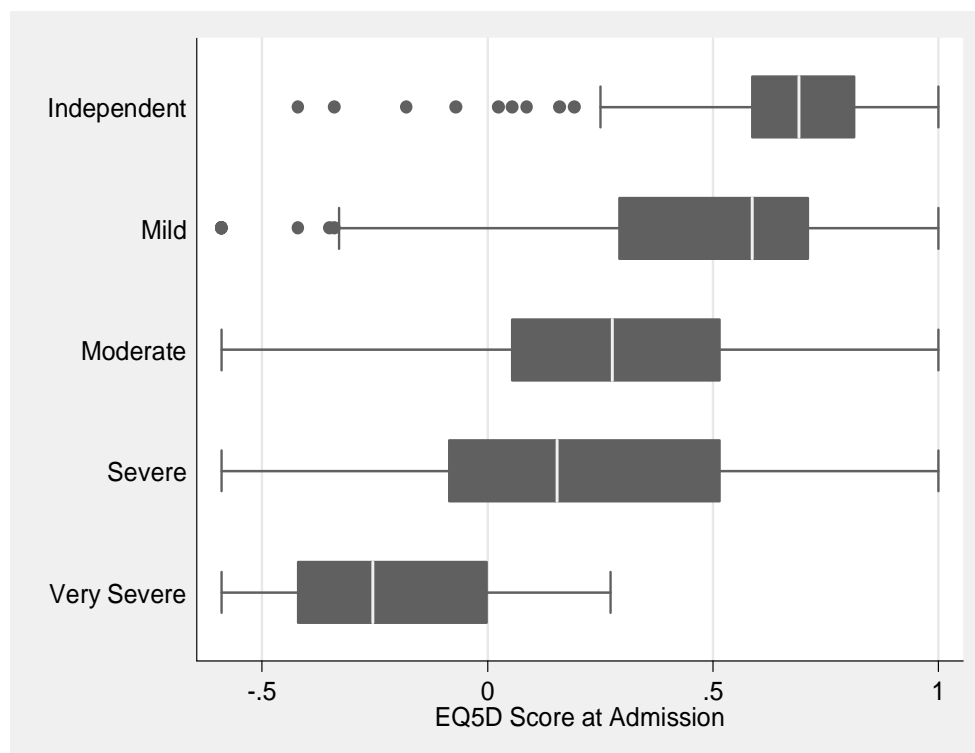


Figure 9.2: Box-Plot: Barthel & EQ-5D at admission



An examination of figures 9.3 to 9.7 shows the relationship between the overall Barthel score and the levels on each of the five EQ-5D dimensions. In all depictions, the positive relationship between the two measures is again evident as low Barthel scores are associated with 'extreme problems' on the EQ-5D dimensions and high Barthel scores are associated with 'No problems'. It is however clear also from the box-plots below that changes in the overall Barthel score between levels of the EQ-5D dimensions are more sensitive for certain dimensions than others. For the mobility and Self-care dimensions for instance, there is an approximately 70% change between 'extreme problems' and 'no problems' (from 10.10 to 17.03 and 10.07 to 17.10, respectively). The corresponding changes for the pain and anxiety dimensions were 23% (12.84 to 15.80) and 20% (13.27 to 15.97), respectively.

Figure 9.3: Barthel versus EQ-5D Mobility dimension: Admission Data

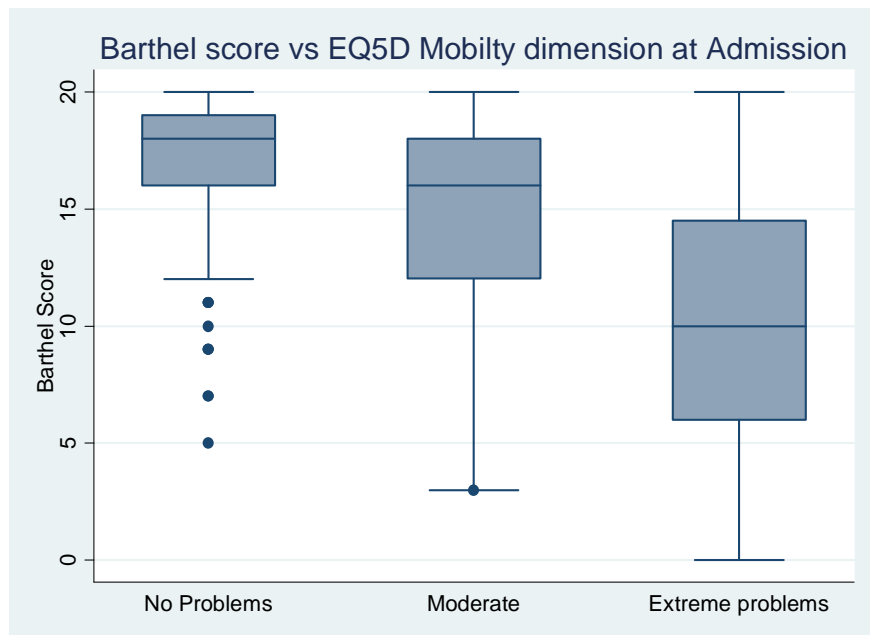


Figure 9.4: Barthel versus EQ-5D Self-Care dimension: Admission Data

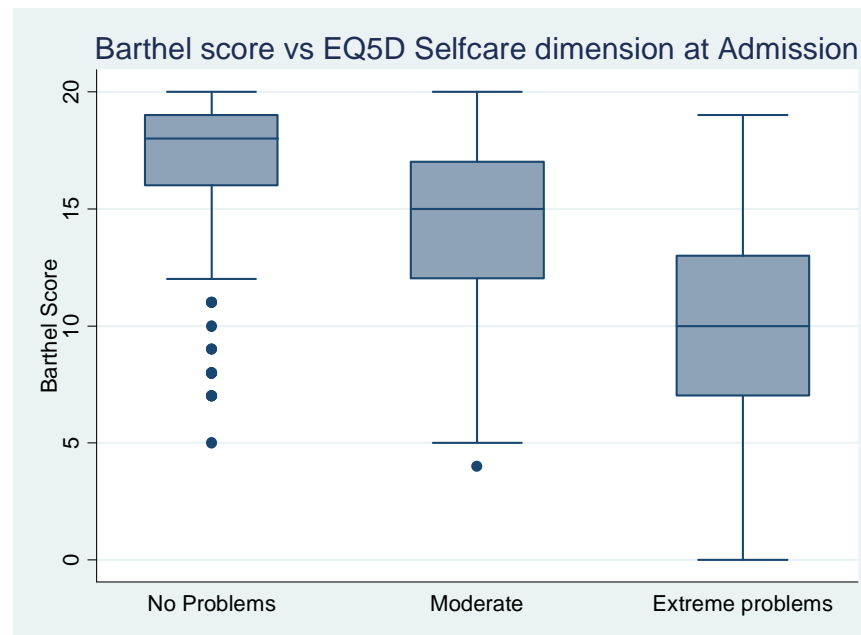


Figure 9.5: Barthel versus EQ-5D Usual activities dimension: Admission Data

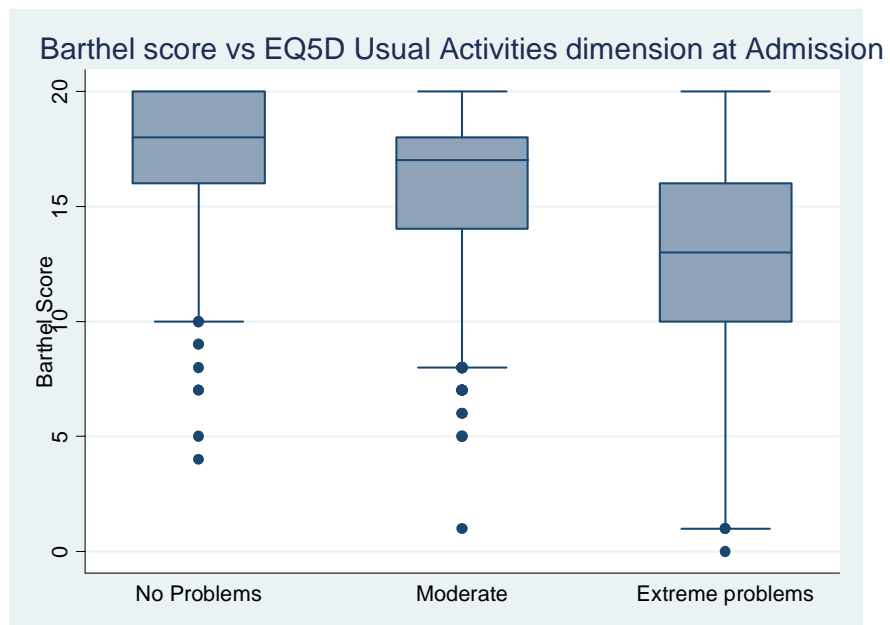


Figure 9.6: Barthel score versus EQ-5D Pain dimension; Admission data

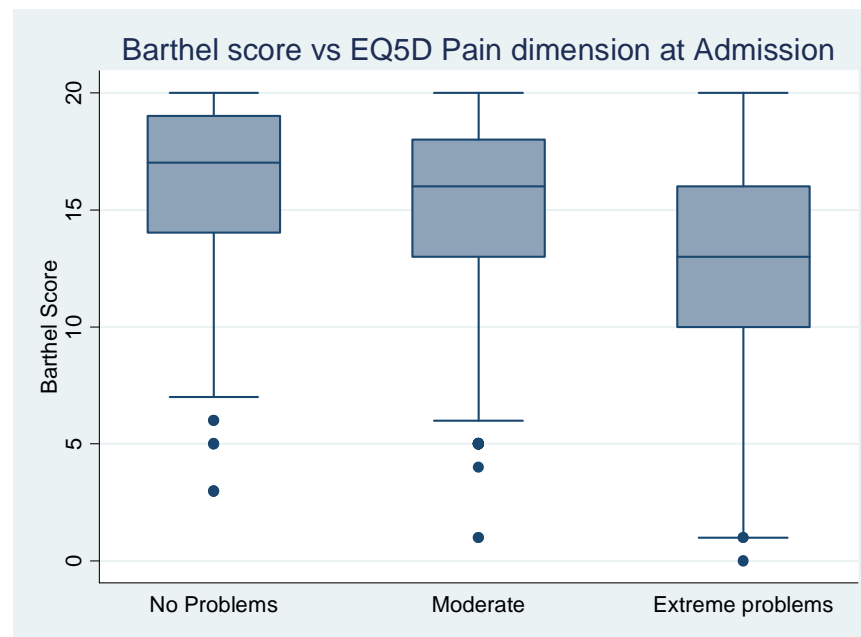
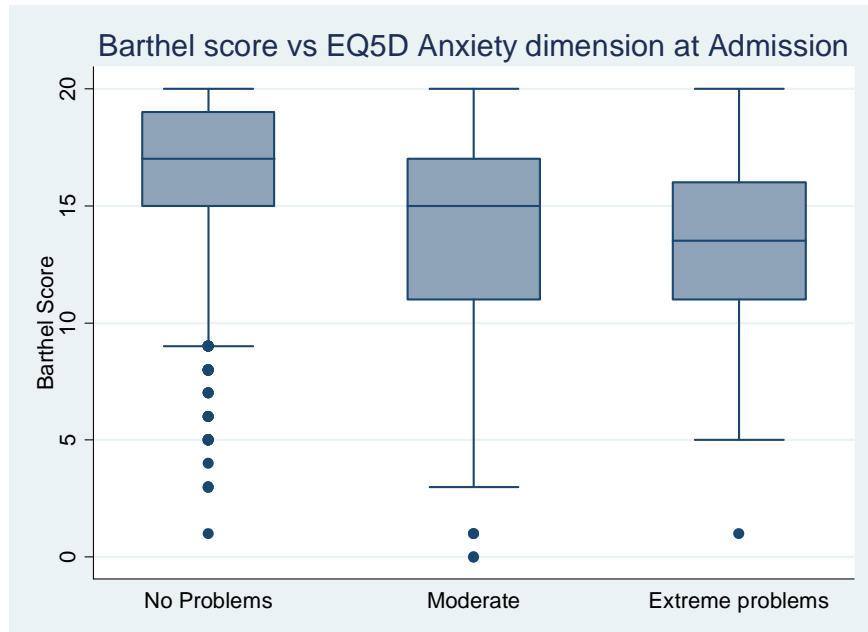


Figure 9.7: Barthel score versus EQ-5D Anxiety dimension; Admission data



9.3.3 Regression results – Predictive ability of the Barthel

Testing the van Exel (2004) model

The van Exel (2004) model was tested on the prediction dataset to see how well it performed. Using this model, a mean predicted EQ-5D tariff score of 0.49 (minimum of -0.25 and maximum of 0.75) was obtained compared to the 0.42 that was obtained for the observed overall EQ-5D tariff score. The RMSE score for the model was 0.330 while the MAE was 0.246. The results affirmed the association of low (high) Barthel scores with very low (high) EQ-5D tariff scores which is an indicator of comparable sensitivity of the two measures.

Tables 9.2 to 9.4 show the results of the 11 regression models described in section 9.2.2 and based on the main dataset, the admission dataset. Further, only CLAD results are shown in this chapter for models 1 to 9. The CLAD results for the split admission datasets are presented in Tables A2 and A3 of the Appendix while the OLS results for both the entire and split admission datasets are shown in tables A5 to A8 of the Appendix.

Model 1

Table 9.2 shows a positive relationship between the overall EQ-5D tariff score and the Barthel overall score (a positive sign for the overall Barthel score coefficient). Age and gender were not significantly related to the overall EQ-5D tariff score. The coefficient of determination (R-squared) was 0.143.

Model 2

This model was the same as model (1) with the exception that age and gender were removed as independent variables. Here again as shown in Table 9.2, the overall Barthel score was positively related to the overall EQ-5D tariff score and the R-squared was 0.140.

Model 3

As shown in Table 9.2, only four Barthel dimensions were statistically significant in this model. These were 'Toilet', 'Dressing', 'Stairs' and 'Bladder'. All except 'Bladder' were positively related to the overall EQ-5D tariff score. Age and gender were again not statistically significant in this model. The R-squared statistic was higher than those for the first two models at 0.168.

Model 4

When age and gender were excluded from model (3), the same covariates that were statistically significant in the model are again significant in this model. These were 'Toilet', 'Dressing', 'Stairs' and 'Bladder'. These covariates also had the same signs as before and the coefficient sizes were also similar but the coefficient of determination (R-squared) was slightly lower at 0.165.

Model 5

In this model, stepwise regression was used to reduce the number of covariates used in model (4). This resulted in only four variables being included in the model: 'Transfer', 'Mobility', 'Dressing' and 'Stairs'. All four covariates were positively

related to the overall EQ-5D tariff score as indicated in Table 9.2. The coefficient of determination was 0.161.

Model 6

Interaction terms were added to model (5) and the result was model (6). None of the main effects and interaction terms were statistically significant but the R-square was 0.168.

Model 7

In table 9.3, eight levels of Barthel dimensions are shown to have been statistically significant in explaining the overall EQ-5D tariff score. These were the 'Independent' level for the 'Toilet' dimension; 'Major help', 'Minor help' and 'Independent' levels for the 'Transfer' dimension; 'Needs help' and 'Independent' for the 'Dressing' dimension and 'Needs help' and 'Independent' levels for the 'Stairs' dimension. The reference category for all of the categorical explanatory variables was the 'dependent/unable' category. All of the statistically significant variables had a positive relationship with EQ-5D. The R-squared statistic was 0.179.

Model 8

When age and gender were removed from the preceding model, all the variables that were significant in that model were still significant except the 'Independent' level for the 'Toilet' dimension was no longer statistically significant. The 'Major help', 'Minor help' and 'Independent' levels for the 'Transfer' dimension; 'Needs help' and 'Independent' for the 'Dressing' dimension and 'Needs help' and 'Independent' levels

for the ‘Stairs’ dimension were still statistically significant. The R-squared was slightly lower though at 0.176.

Model 9

This was a reduced version of Model (8) where stepwise regression was used to determine the variables to be included as covariates. These are shown in Table 9.3. Here, all the levels which were significant in model (8) were again significant in this model. However, the ‘Needs help’ and ‘Independent’ for the ‘Mobility’ dimension which were previously not statistically significant were now statistically significant. An R-squared value of 0.168 was obtained.

Model 10

Table 9.4 shows the results of the multinomial logistic regression results. While the coefficients in this model are not easy to interpret (Greene 1997), what they showed was that some EQ-5D dimensions were more likely to be associated with particular Barthel dimensions. The EQ-5D ‘Mobility’ dimension was likely to be associated with the ‘Transfer’, ‘Mobility’, ‘Stairs’ and ‘Bowels’ dimensions. The Barthel dimensions that were associated with the ‘Self-Care’ domain of the EQ-5D were ‘Transfer’, ‘Dressing’, ‘Stairs’ and ‘Bathing’. The ‘Toilet’, ‘Dressing’, ‘Stairs’, ‘Bathing’ and ‘Bladder’ dimensions of the Barthel index were associated with the ‘Usual Activities’ domain of the EQ-5D. Three Barthel dimensions (Transfer, Stairs and Bladder) were significantly associated with the ‘Pain/Discomfort’ domain while only the ‘Feeding’ Barthel dimension was statistically significant in explaining the ‘Anxiety/Depression’ domain. Nearly all of the statistically significant coefficients had negative signs. The only exceptions were ‘Toilet’ and ‘Bladder’ Barthel

dimensions in the 'Usual activities' multinomial regression model; the 'Bladder' Barthel dimension in the Pain/Discomfort model and the 'Feeding' Barthel dimension in the Anxiety/Depression model. The R-squared values with associated models in parentheses were 0.145 (Mobility), 0.194 (Self-Care), 0.140 (Usual Activities), 0.045 (Pain/Discomfort) and 0.052 (Anxiety/Depression).

Model 11

The results also presented in table 9.4 show that the overall Barthel score was statistically related to all EQ-5D dimensions. All coefficients were negative and the R-squared values with associated models in parentheses were 0.104 (Mobility), 0.152 (Self-Care), 0.083 (Usual Activities), 0.027 (Pain/Discomfort) and 0.038 (Anxiety/Depression).

Table 9.2: CLAD model – Barthel dimensions entered as continuous variables: Admission Data

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Variables	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Total Barthel score	0.048*	0.004	0.048*	0.003								
Age	0.014	0.013			0.016	0.014						
Gender (1 = female, 0 = male)	-0.031	0.019			-0.026	0.023						
Barthel – Grooming					0.004	0.037	-0.000	0.041				
Barthel – Toilet					0.097*	0.041	0.100*	0.039				
Barthel – Feeding					0.013	0.037	0.000	0.042				
Barthel – Transfer					0.074	0.047	0.068	0.052	0.098*	0.036	0.056	0.094
Barthel – Mobility					0.044	0.037	0.046	0.042	0.081*	0.029	0.086	0.085
Barthel – Dressing					0.096*	0.021	0.104*	0.024	0.105*	0.154	0.051	0.085
Barthel – Stairs					0.055*	0.013	0.052*	0.012	0.052*	0.009	0.091	0.091
Barthel – Bathing					0.013	0.026	-0.000	0.025				
Barthel – Bladder					-0.041*	0.019	-0.042*	0.018				
Barthel – Bowels					-0.003	0.041	-0.000	0.045				
Barthel - Transfer x Toilet											0.013	0.041

Barthel - Mobility x Toilet											0.030	0.040
Barthel - Dressing x Toilet											0.011	0.038
Barthel - Stairs x Toilet											-0.041	0.044
Constant	-0.255*	0.078	-0.240*	0.060	-0.129	0.113	-0.076	0.110	-0.16*	0.058	-0.102	0.070
R-Squared		0.143		0.140		0.168		0.165		0.161		0.168

Dependent Variable; EQ-5D * Statistically significant at 5% level

Table 9.3: CLAD model – Barthel dimensions entered as categorical variables: Admission Data

	Model 7		Model 8		Model 9	
Variables	Coeff.	SE	Coeff.	SE	Coeff.	SE
Age	0.018	0.012				
Gender (1 = female, 0 = male)	-0.034	0.022				
Grooming – independent	-0.016	0.041	0.000	0.044		
Toilet – needs help	0.051	0.089	0.037	0.087		
Toilet – independent	0.184*	0.086	0.174	0.095		
Toilet – dependent (Reference category)						
Feeding – needs help	-0.154	0.142	-0.123	0.180		
Feeding – independent	-0.139	0.143	-0.123	0.174		
Feeding – unable (Reference category)						
Transfer – major help	0.481*	0.130	0.475*	0.132	0.446*	0.118
Transfer – minor help	0.436*	0.131	0.442*	0.130	0.394*	0.110
Transfer – independent	0.489*	0.131	0.519*	0.126	0.465*	0.112
Transfer – unable (Reference category)						
Mobility – wheelchair independent	-0.009	0.182	-0.002	0.174	-0.027	0.174
Mobility – needs help	0.128	0.098	0.110	0.104	0.175*	0.075
Mobility – independent	0.156	0.097	0.126	0.110	0.275*	0.079
Mobility – immobile (Reference category)						
Dressing – needs help	0.147*	0.054	0.144*	0.059	0.162*	0.058
Dressing – independent	0.243*	0.059	0.242*	0.070	0.266*	0.058
Dressing – dependent (Reference category)						
Stairs – needs help	0.085*	0.035	0.098*	0.031	0.104*	0.034
Stairs – independent	0.099*	0.028	0.098*	0.024	0.104*	0.027

Stairs – unable (Reference category)						
Bathing – independent	0.026	0.029	0.006	0.023		
Bladder – occasional accident	-0.031	0.048	-0.033	0.059		
Bladder – continent	-0.069	0.041	-0.069	0.050		
Bladder – incontinent (Reference category)						
Bowels – occasional accident	-0.046	0.115	-0.044	0.150		
Bowels – continent	-0.028	0.116	-0.028	0.141		
Bowels – incontinent (Reference category)						
Constant	-0.272	0.128	-0.252	0.211	- 0.419*	0.100
R-Squared		0.179		0.176		0.168

Dependent Variable; EQ-5D * Statistically significant at 5% level

Table 9.4: Multinomial Logit using Barthel dimensions as independent variables: Admission Data (Models 10 & 11)

	Mobility				Self-Care				Usual Activities				Pain/Discomfort				Anxiety/Depression			
	Some Problems		Extreme Problems		Some Problems		Extreme Problems		Some Problems		Extreme Problems		Some Problems		Extreme Problems		Some Problems		Extreme Problems	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Grooming	0.283	0.258	0.692	0.433	-0.357	0.220	-0.291	0.339	-0.357	0.316	-0.031	0.336	-0.198	0.201	0.038	0.280	-0.194	0.190	-0.242	0.289
Toilet	-0.409	0.314	-0.250	0.426	-0.055	0.243	-0.311	0.315	0.711*	0.295	0.341	0.304	-0.157	0.206	-0.061	0.268	-0.161	0.187	-0.225	0.274
Feeding	-0.137	0.281	-0.011	0.400	-0.257	0.243	-0.620	0.317	-0.075	0.324	-0.344	0.335	0.058	0.200	0.065	0.263	0.257	0.189	0.766*	0.300
Transfer	-0.250	0.283	-0.913*	0.367	-0.595*	0.237	-0.661*	0.297	-0.289	0.311	-0.528	0.320	-0.033	0.191	-0.624*	0.229	-0.328	0.170	-0.343	0.243
Mobility	-0.534	0.299	-1.249*	0.343	0.236	0.179	0.065	0.227	-0.403	0.283	-0.396	0.290	-0.063	0.161	-0.308	0.189	-0.136	0.141	-0.205	0.192
Dressing	0.094	0.207	0.113	0.346	-0.781*	0.181	-2.226*	0.285	-0.564*	0.249	-1.079*	0.266	-0.024	0.161	-0.062	0.225	-0.136	0.152	-0.196	0.235
Stairs	-0.545*	0.121	-0.906*	0.292	-0.195*	0.099	-0.606*	0.227	-0.063	0.129	-0.449*	0.154	-0.222*	0.099	-0.289	0.154	0.035	0.096	-0.322	0.173
Bathing	-0.117	0.232	-1.166	0.792	-0.359	0.193	-2.304*	1.089	-0.308	0.243	-1.063*	0.339	0.261	0.203	-0.172	0.344	0.015	0.198	-0.488	0.405
Bladder	0.162	0.196	0.479	0.320	0.200	0.166	0.351	0.244	0.461*	0.202	0.440*	0.222	0.333*	0.146	0.441*	0.210	-0.176	0.140	0.023	0.222
Bowels	-0.250	0.355	-0.918*	0.444	0.038	0.254	-0.236	0.332	0.161	0.294	0.309	0.316	-0.293	0.223	-0.118	0.295	-0.296	0.205	-0.197	0.305
Constant	5.170*	0.966	5.894*	1.112	3.452*	0.689	5.146*	0.824	2.433*	0.860	4.006*	0.877	1.313*	0.524	1.341*	0.642	1.839*	0.475	-0.258	0.707
Total Barthel	-0.196*	0.032	-0.452*	0.044	-0.241*	0.027	-0.526*	0.038	-0.110*	0.034	-0.310*	0.036	-0.053*	0.020	-0.178*	0.026	-0.129*	0.019	-0.172*	0.027
Constant	4.756	0.544	5.584	0.641	4.486	0.458	6.537	0.551	3.257	0.576	5.519	0.592	1.447	0.325	1.897	0.382	1.669	0.291	0.939	0.395

* Statistically significant at 5% level

9.3.4 Predicted EQ-5D Tariff Scores and Goodness-of-fit – Admissions Data

Table 9.5 presents summary results of the key goodness-of-fit statistics from the 11 models. A look at the results from the ‘within-sample predictions’ shows that the observed mean EQ-5D tariff score was 0.42. These results also indicate that all models had a predicted mean EQ-5D tariff score that was greater than the observed EQ-5D tariff score. The predicted EQ-5D tariff scores ranged from 0.44 to 0.50 i.e. within about 4 to 19 percentage points of the true EQ-5D tariff score. Models (8) and (11*) had predicted EQ-5D tariff scores closest to the observed score while models (10#) and (11#) had the farthest. The observed EQ-5D tariff score had a ‘full’ range (-0.59 to 1.00). None of the models produced predicted values with such a wide range. The widest range for predicted EQ-5D scores amongst the models was for models (7) and (11*) followed by model (10*) and the smallest range was for model (5). The van Exel et al (2004) model had the third highest predicted EQ-5D tariff score (0.49) and the range for this predicted score was from -0.25 to 0.75. In terms of correlation, the results indicate that models (8) and (9) predicted EQ-5D tariff scores that had the highest correlation with the observed EQ-5D tariff score (0.488). Other fairly high correlation figures were obtained for models (7), (5) and (4). The lowest correlation was found in the four multinomial logistic regression models. Again, the van Exel model did not perform very well as its correlation value was among the lowest, higher than only those for models (10*), (10#), (11*) and (11#). The RMSEs ranged from 0.324 to 0.361 while the MAEs varied from 0.240 to 0.277. In order of performance measured by the combined RMSE and MAE scores, the results indicate that models (3), (7), (4), (6) and (8) had lower RMSE and MAE scores than other models. More weight was placed on the RMSEs in determining performance based on the observations of Hynman and Koehler (2006).

The relevant equations for these five models were:

$$\begin{aligned} \text{EQ-5D} = & -0.129 + 0.016\text{Age} - 0.026\text{Gender} + 0.004\text{Grooming} + 0.097\text{Toilet} + 0.013\text{Feeding} \\ & + 0.074\text{Transfer} + 0.044\text{Mobility} + 0.096\text{Dressing} + 0.055\text{Stairs} + 0.013\text{Bathing} - \\ & 0.041\text{Bladder} - 0.003\text{Bowels} \end{aligned} \quad (\text{Model 3})$$

$$\begin{aligned} \text{EQ-5D} = & -0.076 - 0.000\text{Grooming} + 0.100\text{Toilet} + 0.000\text{Feeding} + 0.068\text{Transfer} + \\ & 0.046\text{Mobility} + 0.104\text{Dressing} + 0.052\text{Stairs} - 0.000\text{Bathing} - 0.042\text{Bladder} \\ & - 0.000\text{Bowels} \end{aligned} \quad (\text{Model 4})$$

$$\begin{aligned} \text{EQ-5D} = & -0.102 + 0.056\text{Transfer} + 0.086\text{Mobility} + 0.051\text{Dressing} + 0.091\text{Stairs} + \\ & 0.013\text{Transfer} \times \text{Toilet} + 0.030\text{Mobility} \times \text{Toilet} + 0.011\text{Dressing} \times \text{Toilet} - 0.041\text{Stairs} \times \\ & \text{Toilet} \end{aligned} \quad (\text{Model 6})$$

$$\begin{aligned} \text{EQ-5D} = & -0.272 + 0.018\text{Age} - 0.034\text{Gender} - 0.016\text{Grooming_Independent} + \\ & 0.051\text{Toilet_NeedsHelp} + 0.184\text{Toilet_Independent} - 0.154\text{Feeding_NeedsHelp} - \\ & 0.139\text{Feeding_Independent} + 0.481\text{Transfer_MajorHelp} + 0.436\text{Transfer_MinorHelp} + \\ & 0.489\text{Transfer_Independent} - 0.009\text{Mobility_WheelchairIndependent} + \\ & 0.128\text{Mobility_NeedsHelp} + 0.156\text{Mobility_Independent} + 0.147\text{Dressing_NeedsHelp} + \\ & 0.243\text{Dressing_Independent} + 0.085\text{Stairs_NeedsHelp} + 0.099\text{Stairs_Independent} + \\ & 0.026\text{Bathing_Independent} - 0.031\text{Bladder_OccasionalAccident} - 0.069\text{Bladder_Continent} - \\ & 0.046\text{Bowels_OccasionalAccident} - 0.028\text{Bowels_Continent} \end{aligned} \quad (\text{Model 7})$$

$$\begin{aligned} \text{EQ-5D} = & -0.252 + 0.000\text{Grooming_Independent} + 0.037\text{Toilet_NeedsHelp} + \\ & 0.174\text{Toilet_Independent} - 0.123\text{Feeding_NeedsHelp} - 0.123\text{Feeding_Independent} + \\ & 0.475\text{Transfer_MajorHelp} + 0.442\text{Transfer_MinorHelp} + 0.519\text{Transfer_Independent} - \\ & 0.002\text{Mobility_WheelchairIndependent} + 0.110\text{Mobility_NeedsHelp} + \\ & 0.126\text{Mobility_Independent} + 0.144\text{Dressing_NeedsHelp} + 0.242\text{Dressing_Independent} + \\ & 0.098\text{Stairs_NeedsHelp} + 0.098\text{Stairs_Independent} + 0.006\text{Bathing_Independent} - \\ & 0.033\text{Bladder_OccasionalAccident} - 0.069\text{Bladder_Continent} - \\ & 0.044\text{Bowels_OccasionalAccident} - 0.028\text{Bowels_Continent} \end{aligned} \quad (\text{Model 8})$$

The multinomial logistic models had the largest RMSEs. Models (10*) and (11*) had RMSEs of 0.353 and 0.361, respectively, which were higher than those for models (10#) and (11#). Models (10*) and (11*) also had the largest MAEs (0.270 and 0.277, respectively). The RMSE and MAE values for the van Exel model were fairly high at 0.330 and 0.246, respectively.

9.3.5 Robustness - Out-sample Prediction – Discharge Data

Table 9.5 also presents the summary statistics from the validation dataset (discharge data). In this validation exercise, the models estimated from the admissions data were run on the discharge data and predictions obtained. The mean EQ-5D tariff index for the discharge data was 0.61. With the exception of models (3) to (6), all predicted mean EQ-5D tariff scores were below 0.61. The range for the predicted EQ-5D tariff scores (maximum predicted EQ-5D score – minimum predicted EQ-5D score) that were greater than the observed EQ-5D tariff score was from 0.85 to 1.01 while that for predicted EQ-5D tariff scores that were smaller than the observed values was from 0.96 to 1.35. Predicted EQ-5D tariff values from the van Exel et al model and model (10#) were closest to the observed EQ-5D tariff scores while those from models (3), (4) and (5) were the farthest. As in the admission dataset, the observed EQ-5D tariff score had a ‘complete’ range (-0.59 to 1.00). None of the models produced predicted values with such a wide range. The widest range for predicted EQ-5D scores amongst the models was for models (7), (8), (10*) and (11*) while the narrowest range was again for model (5) which was matched by that of model (6). The highest correlation figures were obtained for the van Exel et al model and model (2) but other models that had high correlation values in the admission dataset also had fairly high correlations here too. The lowest correlation was again for EQ-5D scores predicted by the four multinomial logistic regression models. On average, the RMSEs and MAEs were much smaller in this dataset. The RMSEs ranged from 0.251 to 0.367 while the MAEs vary from 0.192 to 0.267. When the combined RMSE and MAE scores are considered, the results indicate that models (1), (9), (2), (7) and (8) had lower RMSE and MAE scores than other models. The van Exel et al (2004) model also had fairly low RMSE and MAE scores. While the multinomial logistic models were not the best

performers, the results show that models (3), (4) and (5) had the largest RMSEs (MAEs): 0.327 (0.225), 0.327 (0.224) and 0.367 (0.267), respectively. Models (10#) and (11#) performed better than (10*) and (11*).

9.3.6 Robustness – Using split Admission Prediction dataset

In Table 9.6, the goodness-of-fit results of running the models described in section 9.2.2 on a split sample of the admission dataset are shown. The regression results were broadly similar to those reported for the entire admissions data in section 9.3.3 and are presented in the A7 of the appendix. The discussion in this section is based on the goodness-of-fit results only. Using the prediction dataset, the observed mean EQ-5D tariff score was found to be 0.43 which is close to that for the entire admissions data (0.42). As was the case for the entire admissions dataset, all models had predicted mean EQ-5D tariff indices that were greater than this observed EQ-5D tariff. The predicted EQ-5D tariff scores ranged from within about 5 to 21 percentage points of the true tariff score. The predicted EQ-5D scores closest to the observed EQ-5D tariff score were from models (10*) and (11*) but another fairly close value was predicted by model (6). Models (1), (10#) and (11#) predicted values that were furthest away from the observed EQ-5D tariff value. The van Exel et al (2004) model had a fairly high predicted EQ-5D tariff score which was higher than those predicted by all but three models. The observed EQ-5D tariff score had a full range (-0.59 to 1.00). Though none of the models produced predicted values with such a wide range, model (8) produced the widest range followed by models (10*) and (11*). The narrowest range was again for model (5). In terms of correlation, the results indicate that models (3), (7), (4) and (5) predicted EQ-5D tariff scores that had the highest correlation with the observed EQ-5D tariff score. The lowest again were for the multinomial logistic

models. The van Exel et al (2004) model had a fairly low correlation figure. The RMSEs ranged from 0.326 to 0.378 while the MAEs varied from 0.240 to 0.290. The models that have a lower combined RMSE and MSE were (3), (4), (5), (6) and (7) while the multinomial logistic models had the largest RMSEs (0.350 for model (10*), 0.354 for model (11*), 0.369 for model (10#) and 0.378 for model (11#). Models (10#) and (11#) had the largest MAEs. The van Exel et al (2004) model had the seventh highest RMSE score and the tenth highest MAE score among all the models.

9.3.7 Robustness – Using split Admission Validation dataset

This section reports the results of running the models estimated using the split prediction dataset (whose results have been reported in section 9.3.6 and A7 in the appendix) on the split admissions validation dataset. The summary statistics from this exercise are also reported in Table 9.6. The mean EQ-5D tariff index for this validation dataset was 0.42. All predicted mean EQ-5D tariff scores were above 0.42 and ranged between 5 and 21 percentage points of the observed EQ-5D score. The predicted EQ-5D scores closest to the observed EQ-5D tariff score were from models (6) and (7) while other fairly close values were predicted by models (2), (3) and (4), and (5), (8) and (11*). The predicted values that were furthest away from the observed EQ-5D tariff value were again from models (10#) and (11#). The van Exel et al (2004) model had a fairly high predicted EQ-5D tariff score which was higher than those predicted by all but five models. The observed EQ-5D tariff score had a ‘complete’ range (-0.59 to 1.00). Though none of the models produced predicted values with such a wide range, models (10*) and (7) produced the widest range followed by models (11*) and (10#). The narrowest range was again for model (5). In terms of correlation, the results indicate that models (7), (9), (5) and (6) predicted EQ-

5D tariff scores that had the highest correlation with the observed EQ-5D tariff score. The lowest were again for the multinomial logistic models. The van Exel et al (2004) model had a fairly high correlation figure. The RMSEs ranged from 0.307 to 0.345 while the MAEs vary from 0.201 to 0.261. The models that had a lower combined RMSE and MSE were (2), (3), (4) and (7) while the multinomial logistic models had the largest RMSEs (0.339 for model (10*), 0.343 for model (11*), 0.345 for model (10#) and 0.341 for model (11#). Models (10*) and (11*) had the largest MAEs. The van Exel et al (2004) model had the ninth largest RMSE score and the joint second smallest MAE score among all the models.

A summary of the results presented in sections 9.3.4 to 9.3.7 is presented in Table 9.7.

9.3.8 Interpreting the goodness-of-fit results

From the results of the goodness-of-fit analysis, it was difficult to find a consistent message coming from the analyses of all the datasets. However, some fairly reliable messages seem to come through from the analyses considered. It is also important to point out here that in interpreting the results of the goodness-of-fit, more weight was placed on the results reported in sections 9.3.4, 9.3.6 and 9.3.7. This is because, and as pointed out in section 9.2.4, the results presented in section 9.3.5 were based on a validation dataset (the discharge dataset) which could not be strictly considered to be independent. The results indicated that in order of performance measured by the combined RMSE and MAE scores models (3), (7), (4) and (6) performed better than the other models. All models except (7) treated the Barthel index dimensions as continuous variables. The RMSEs and MAEs were lower for these models. This means that these models had the lowest error when predicting the mean EQ-5D tariff scores. The error was a measure of how close or far from the actual EQ-5D tariff

score the predicted scores were. In interpreting the result of the RMSE and MAE, more weight was placed on the RMSE because it is considered to be more robust than the MAE especially in the face of outliers (Hyndman & Koehler 2006). But as pointed out by Bansback et al (2007), there is no definition of the level of the RMSE that should be considered acceptable for fitting purposes and the only judgement that can be made is one of relative performance of models. But based on the principle of parsimony, models (3) and (4) would be preferred choices while model (4) would be ideal when one does not have to rely on patient characteristics to fit the model. The results of the RMSE and MAE are borne out by the correlation figures as well; the four models had the highest or fairly high correlation to the actual EQ-5D tariff score compared to other models. Another feature of these models is that they had a fairly wide range of predicted scores compared to the others which result may indicate there was not much of a ceiling effect problem with these models. The multinomial logistic regression models performed the worst in terms of predicting the EQ-5D tariff scores.

Table 9.5: Predicted versus Actual EQ-5D utility scores: Admission and Discharge Datasets – CLAD Results

	Within-Sample Prediction: Admission Data (Prediction dataset)						Out-of-Sample Prediction: Discharge Data (Validation dataset)					
	Mean	Max	Min	Correlation	RMSE	MAE	Mean	Max	Min	Correlation	RMSE	MAE
Observed	0.42	1.00	-0.59	-	-	-	0.61	1.00	-0.59	-	-	-
vanExel Model	0.49	0.75	-0.25	0.465	0.330	0.246	0.59	0.75	-0.25	0.475	0.279	0.202
Model (1)	0.46	0.73	-0.28	0.467	0.325	0.245	0.57	0.76	-0.24	0.434	0.251	0.192
Model (2)	0.46	0.73	-0.31	0.465	0.325	0.245	0.56	0.72	-0.24	0.475	0.280	0.208
Model (3)	0.48	0.79	-0.21	0.477	0.324	0.240	0.76	1.00	-0.01	0.473	0.327	0.225
Model (4)	0.45	0.80	-0.34	0.481	0.325	0.241	0.76	0.97	0.03	0.474	0.327	0.224
Model (5)	0.48	0.71	-0.16	0.483	0.326	0.243	0.83	0.96	0.11	0.474	0.367	0.267
Model (6)	0.46	0.78	-0.11	0.478	0.325	0.240	0.65	0.75	-0.10	0.442	0.287	0.199
Model (7)	0.47	1.00	-0.45	0.483	0.324	0.243	0.57	0.89	-0.49	0.471	0.280	0.207
Model (8)	0.44	0.89	-0.52	0.488	0.325	0.240	0.57	0.84	-0.47	0.469	0.280	0.205
Model (9)	0.48	0.69	-0.42	0.488	0.326	0.240	0.57	0.69	-0.42	0.469	0.277	0.203
Model (10*)	0.45	0.96	-0.46	0.380	0.353	0.270	0.53	0.96	-0.36	0.345	0.317	0.245
Model (11*)	0.44	1.00	-0.45	0.341	0.361	0.277	0.52	1.00	-0.35	0.322	0.326	0.252
Model (10#)	0.50	0.69	-0.42	0.454	0.340	0.240	0.58	0.81	-0.42	0.408	0.291	0.206
Model (11#)	0.50	0.81	-0.42	0.405	0.347	0.251	0.57	0.69	-0.42	0.391	0.294	0.212

*Denotes results obtained after Monte Carlo simulations

Denotes results obtained using the 'indirect method'

Table 9.6: Predicted versus Actual EQ-5D utility scores: Split Admission dataset samples – CLAD Results

	Within-Sample Prediction: Prediction Dataset						Out-of-Sample Prediction: Validation dataset					
	Mean	Max	Min	Correlation	RMSE	MAE	Mean	Max	Min	Correlation	RMSE	MAE
Observed	0.43	1.00	-0.59	-	-	-	0.42	1.00	-0.59	-	-	-
vanExel Model	0.50	0.75	-0.25	0.448	0.334	0.248	0.47	0.75	-0.20	0.510	0.317	0.238
Model (1)	0.51	0.80	-0.14	0.455	0.335	0.248	0.48	0.75	-0.14	0.372	0.323	0.239
Model (2)	0.49	0.72	-0.18	0.448	0.332	0.247	0.46	0.71	-0.16	0.510	0.313	0.238
Model (3)	0.47	0.78	-0.23	0.468	0.326	0.241	0.46	0.78	-0.23	0.527	0.312	0.238
Model (4)	0.47	0.86	-0.26	0.463	0.327	0.241	0.46	0.79	-0.24	0.527	0.311	0.238
Model (5)	0.48	0.69	-0.20	0.465	0.327	0.242	0.46	0.69	-0.11	0.534	0.315	0.242
Model (6)	0.46	0.98	-0.23	0.458	0.328	0.241	0.45	0.69	-0.08	0.534	0.314	0.241
Model (7)	0.49	0.85	-0.16	0.473	0.329	0.240	0.44	0.74	-0.53	0.543	0.307	0.235
Model (8)	0.49	1.00	-0.56	0.447	0.335	0.241	0.46	0.76	-0.29	0.529	0.314	0.239
Model (9)	0.48	0.69	-0.30	0.448	0.334	0.241	0.48	0.69	-0.16	0.539	0.320	0.242
Model (10*)	0.45	0.95	-0.52	0.382	0.350	0.244	0.47	0.94	-0.38	0.387	0.339	0.256
Model (11*)	0.45	1.00	-0.45	0.347	0.354	0.254	0.46	0.92	-0.32	0.365	0.343	0.261
Model (10#)	0.52	0.85	-0.42	0.409	0.369	0.281	0.50	0.81	-0.42	0.364	0.345	0.245
Model (11#)	0.51	0.69	-0.42	0.372	0.378	0.290	0.51	0.69	-0.42	0.360	0.341	0.249

*Denotes results obtained after Monte Carlo simulations

Denotes results obtained using the 'indirect method'

Table 9.7: Summary of performance of predicted EQ-5D scores[§] - CLAD Models

Characteristic of predicted EQ-5D score:	Models based on Admissions data (Prediction dataset) n = 1189	Models based on Discharge data (Validation dataset) n = 910	Models based on Split Admissions dataset (Prediction dataset) n = 793	Models based on Split Admissions dataset (Validation dataset) n = 396
Value close to observed EQ-5D	8, 11*	Van Exel, 10#	3, 4, 6, 10*, 11*,	2, 3, 4, 5, 6, 7, 8, 11*
Value further from observed EQ-5D	10#, 11#	3, 4, 5	1, 10#, 11#	10#, 11#
Wide range [£]	7, 11*, 10*	7, 8, 10*, 11*	8, 10*, 11*	11*, 10#
Narrow range [£]	5	5, 6	5	5
High correlation to observed EQ-5D	4, 5, 8, 9, 7	van Exel, 2, 3, 4, 5, 7	3, 4, 5, 7	5, 6, 7, 9
Low correlation to observed EQ-5D	10*, 11*, 10#, 11#	10*, 11*, 10#, 11#	10*, 11*, 10#, 11#	10*, 11*, 10#, 11#
Small RMSE	van Exel, 1, 2, 3, 4, 6, 7, 8	van Exel, 1, 2, 7, 8, 9	3, 4, 5, 6, 7	2, 3, 4, 6, 7, 8
Small MAE	3, 4, 6, 8, 9, 10#	van Exel, 1, 2, 6, 7, 8, 9	3, 4, 5, 6, 7, 8, 9	van Exel, 2, 3, 4, 6, 7, 8
Large RMSE	10*, 11*, 10#, 11#	3, 4, 5, 11*	10*, 11*, 10#, 11#	10*, 11*, 10#, 11#
Large MAE	10*, 11*, 11#	5, 10*, 11*	11*, 10#, 11#	10*, 11*, 10#, 11#

[§] The cut-off points for deciding the allocations of models in this table were chosen arbitrary and differed according to dataset. The only judgement that was made was that of relative performance.

[£] Range = Maximum predicted EQ-5D – Minimum predicted EQ-5D

Table 9.8: Comparisons of the mean, median, RMSE and MAE for 4 CLAD Models

		Model (3)	Model (4)	Model (6)	Model (7)
Means	D1	0.480	0.450	0.460	0.470
	D2	0.470	0.470	0.470	0.490
	D3	0.460	0.460	0.460	0.490
Medians	D1	0.530	0.530	0.540	0.550
	D2	0.530	0.530	0.540	0.550
	D3	0.540	0.540	0.520	0.540
RMSEs	D1	0.324	0.325	0.325	0.324
	D2	0.326	0.327	0.328	0.329
	D3	0.312	0.311	0.314	0.317
MAEs	D1	0.240	0.241	0.240	0.243
	D2	0.241	0.241	0.242	0.240
	D3	0.238	0.238	0.241	0.235

D1 - Entire admissions dataset;

D2 – Split Admission Prediction dataset;

D3 – Split Admission Validation dataset

9.4 Performance of the models in the validation datasets

Two validation datasets were used in this chapter. These were the discharge dataset and one-third of the split admission dataset. These two datasets were used to validate the findings from the models used in the main datasets by testing the ability of the models to accurately predict utility scores of “out-of-sample” subjects. To critically evaluate the performance of these validation datasets, the next sections look at discrepancies or variances in the results obtained together with discussions about why these differences may exist and then at the similarities or convergences in the findings from the prediction and validation datasets. Thereafter, the question of whether or not the mapping functions were stable when used in the validation datasets is addressed. The key statistics that will form the basis of this critical evaluation are presented in Table 9.7.

9.4.1 Discrepancies between the prediction and validation datasets

First, the results from the models based on the entire admissions (prediction) dataset are compared to those obtained when these models are run on the discharge (validation) dataset. In terms of the bias of predicted EQ-5D values, the results from prediction dataset show that models (8) and (11) have predicted EQ-5D utility values that are closest to the observed EQ-5D utility scores (within at most 0.02 of the observed EQ-5D utility score) whereas the results from the validation dataset show that model (10#) and the Van Exel model are closest (again within at most 0.02 of the observed EQ-5D utility score). Models (10#) and (11#) are shown to have predicted EQ-5D utility scores furthest from the observed EQ-5D score in the prediction dataset while models (3), (4) and (5) have this characteristic in the validation dataset. In the prediction dataset, models (3) and (4) were found to have been among the models

with low RMSEs (0.324 and 0.325, respectively) but were found to be amongst the models with a large RMSE in the validation dataset.

When the results from the models based on the split admissions prediction dataset are compared, some disparities can also be seen. In terms of the bias of predicted EQ-5D values, the results from prediction dataset show that models (10*) and (11*) have predicted EQ-5D utility values that are closest to the observed EQ-5D utility scores (within at most 0.02 of the observed EQ-5D utility score) whereas the results from the validation dataset show that models (6) and (7) are closest (also within at most 0.02 of the observed EQ-5D utility score).

9.4.2 Similarities between the prediction and validation datasets

When the results from the models based on the entire admissions (prediction) dataset are compared to those obtained when these models are run on the discharge (validation) dataset, some similarities in the results obtained are evident. All of the models that produced predicted EQ-5D values with a wide range (maximum predicted EQ-5D – minimum predicted EQ-5D) in the prediction dataset (Models 7, 11* and 10*) were shown to produce wide ranges in the validation dataset as well. The ranges in the prediction dataset were from 1.42 to 1.45 and from 1.35 to 1.38 in the validation dataset. In addition, model (5) was found to have a narrow range in both datasets (0.87 and 0.85 in the prediction and validation dataset, respectively). Further, models (4), (5) and (7) predicted EQ-5D utility values that had a high correlation to the observed EQ-5D scores in both datasets (all within 0.007 of each other). For both datasets, the multinomial models were shown to produce predicted EQ-5D scores that had the lowest correlation to the observed EQ-5D values (all within 0.046 of each

other). Further, the van Exel model, 1, 2, 7 and 8 were shown to have the smallest RMSEs in both dataset.

When the results from the two randomly split samples from the admission dataset are compared, more similarities can be seen. Models (3), (4), (6) and (11*) were amongst the models with the least bias (all of their predicted EQ-5D values were within 0.04 units of the observed EQ-5D utility value) in both samples while models (10#) and (11#) had the most bias in the two samples (all at least 0.08 units greater than the observed EQ-5D utility score). Model (5) was again found to have the narrowest range (0.89 and 0.80, in the prediction and validation dataset, respectively). Both samples also showed that the predicted EQ-5D utility scores from the multinomial logistic models had the lowest correlation to the observed EQ-5D score with all of them within at most 0.035 of each other. Models (3), (4), (6) and (7) were found to be amongst the models with the lowest RMSEs (all within at most 0.02 of each other) and lowest MAEs (all within at most 0.003 of each other) in both samples. For both samples, the multinomial logistic regression models had the largest RMSEs (all within at most 0.037 of each other) and largest MAEs (all within at most 0.02 of each other).

9.4.3 Are the mapping functions stable?

The discrepancies highlighted in section 9.4.1 above suggest that the mapping function is unstable as the results in the prediction datasets can not be replicated with accuracy in a validation dataset. This is indeed true especially when one considers the disparities between the results based on the entire admission dataset and those based on the discharge dataset. But three issues may explain why the results between these two datasets were different. The first is that the relationship between the EQ-5D and Barthel scores in two datasets is not the same. The change in the mean EQ-5D scores

from admission to discharge was disproportionate to that in the Barthel index over the same period. While there was a difference of over 45% in the observed mean EQ-5D tariff scores between admission and discharge (from 0.42 to 0.61), that for the Barthel score changed by only about 13% (from 14.80 to 16.70). This meant that the relationship between the two outcome measures at admission was ‘distorted’ when they were compared at discharge. This essentially implied that the admission dataset was dissimilar to the discharge dataset. This may therefore suggest that the regression equations estimated in these analyses may only be suitable for predicting EQ-5D from Barthel scores at baseline. Second, the results show that on average, the RMSEs and MAEs for the discharge dataset were lower than those for the admission prediction dataset. A possible explanation for this could be that there was less variability in the discharge dataset which may have led to the predicted scores being closer to the observed scores on average. Third, the discharge dataset may not be truly independent as some of the individuals who were in the admissions dataset may have also been a part of the discharge dataset.

Another very important consideration to take into account is that of who administered the Barthel and EQ-5D data collection questionnaires. As indicated in section 4.5, the majority of clients were incapable of self-completing the EQ-5D questionnaire and so stated their responses to staff that then filled out the questionnaires. All of the Barthel index questionnaires on the other hand were filled out by staff. While some studies have shown that it is possible for self-reported responses not to differ significantly from proxy responses (Hilari et al. 2007; Muus et al. 2009; Ostbye et al. 1997), others have shown that proxy responses can be significantly biased (Long et al. 1998; Novella et al. 2006; Hung et al. 2007). There is therefore a possibility that the

differences in the way the two outcome measures were collected may have affected the stability of the mapping functions even though this risk is reduced as essentially the same staff filled out both questionnaires.

It is therefore undeniable that based on the results shown in Table 9.7, it is plausible to suggest that for most of the models, the mapping functions were unstable. This is because there are differences in the values of key statistics obtained from the prediction and validation datasets. However, a question still remains as to what should be the universally agreed-upon minimal difference between the statistics for that difference to be considered significantly different. There is still debate, for instance, about what the RMSE/MAE cut-off points or levels should be at which a judgement can be made on whether or not a model has a good fit (Bansback et al. 2007). It can not be denied that such decisions are made case by case and differ according to different situations. For differences in the EQ-5D, for instance, 0.03 and 0.05 (Cheung et al. 2009) and 0.074 (Walters and Brazier, 2005) have been suggested as the minimal clinically important difference (MCID). If the minimum value of 0.03 were to be used as the MCID, then all of the predicted mean EQ-5D scores based in the validation dataset can not be viewed to be significantly different from those predicted using the prediction datasets. This is because none of these differences between the two predicted values were greater than 0.03.

Whilst recognizing the above debates, this chapter takes a conservative approach in gauging the stability of mapping functions. This is done by taking the view that when assessing whether or not the mapping functions are stable, all the models show cause for concern with the exception of the best performing models whose results are fairly

consistent across the prediction and validation datasets. In section 9.3.8, models (3), (4), (6) and (7) were gauged to have performed better than the rest in predicting mean EQ-5D utility scores. Table 9.8 compares these 4 models in terms the mean scores, median scores, RMSEs and MAEs of the predicted EQ-5D utility scores. The results from the discharge dataset are not considered due to its unsuitability as discussed above.

In terms of bias, the predicted EQ-5D utility scores from the better performing models were all within 0.02 of each other. These differences are all below the lower value of the minimally important difference (MID) of 0.03 suggested by Dobrez et al (2007). Walters and Brazier (2005) have also suggested 0.081 as the MID for the median EQ-5D. Using this criteria, all of the models predicted EQ-5D utility values whose median scores were below this MID (maximum was 0.02). In terms of the RMSEs, some studies have shown that differences of up to 0.017 in the RMSEs did not lead to conclusions that RMSEs were significantly different (Marshall et al, 2008; Tsuchiya et al, 2002; Gray et al. 2006). None of the CLAD models selected as better performing models produced RMSEs which differed by more than 0.016 and so these RMSE differences were well within the accepted ranges. Some studies have also shown that in terms of the MAEs, differences of up to 0.01 in the MAEs did not lead to conclusions that MAEs were significantly different (Franks et al, 2004; Gray et al, 2006; Barton et al, 2008b). None of the CLAD models selected as better predictive models produced MAEs which differed by more than 0.008 and so these MAE differences were also well within the accepted ranges. This therefore suggests that even though more stable mapping functions could have been desirable, the mapping functions from the CLAD prediction models were reasonably stable.

Table 9.9: Predicted versus Actual EQ-5D utility scores: Admission and Discharge Datasets – OLS Results

	Within-Sample Prediction: Admission Data (Prediction dataset)						Out-of-Sample Prediction: Discharge Data (Validation dataset)					
	Mean	Max	Min	Correlation	RMSE	MAE	Mean	Max	Min	Correlation	RMSE	MAE
Observed	0.42	1.00	-0.59	-	-	-	0.61	1.00	-0.59	-	-	-
vanExel Model	0.49	0.75	-0.25	0.465	0.330	0.246	0.59	0.75	-0.25	0.475	0.279	0.202
Model (1)	0.42	0.69	-0.23	0.470	0.319	0.251	0.57	0.76	-0.24	0.434	0.269	0.222
Model (2)	0.42	0.64	-0.21	0.465	0.325	0.251	0.56	0.72	-0.24	0.475	0.293	0.236
Model (3)	0.42	0.73	-0.16	0.500	0.314	0.245	0.76	1.00	-0.01	0.483	0.290	0.232
Model (4)	0.42	0.73	-0.13	0.496	0.314	0.245	0.76	0.97	0.03	0.488	0.287	0.229
Model (5)	0.42	0.63	-0.11	0.490	0.316	0.249	0.83	0.96	0.11	0.474	0.290	0.234
Model (6)	0.42	0.70	-0.10	0.493	0.316	0.247	0.65	0.75	-0.10	0.400	0.341	0.299
Model (7)	0.42	0.75	-0.36	0.515	0.311	0.243	0.57	0.89	-0.49	0.469	0.291	0.234
Model (8)	0.42	0.76	-0.35	0.511	0.311	0.243	0.57	0.84	-0.47	0.476	0.290	0.233
Model (9)	0.42	0.62	-0.27	0.499	0.315	0.247	0.57	0.69	-0.42	0.466	0.291	0.234
Model (10*)	0.45	0.96	-0.46	0.380	0.353	0.270	0.53	0.96	-0.36	0.345	0.317	0.245
Model (11*)	0.44	1.00	-0.45	0.341	0.361	0.277	0.52	1.00	-0.35	0.322	0.326	0.252
Model (10#)	0.50	0.69	-0.42	0.454	0.340	0.240	0.58	0.81	-0.42	0.408	0.291	0.206
Model (11#)	0.50	0.81	-0.42	0.405	0.347	0.251	0.57	0.69	-0.42	0.391	0.294	0.212

*Denotes results obtained after Monte Carlo simulations

Denotes results obtained using the 'indirect method'

Table 9.10: Predicted versus Actual EQ-5D utility scores: Split Admissions Datasets – OLS Results

	Within-Sample Prediction: Admission Data (Prediction dataset)						Out-of-Sample Prediction: Discharge Data (Validation dataset)					
	Mean	Max	Min	Correlation	RMSE	MAE	Mean	Max	Min	Correlation	RMSE	MAE
Observed	0.42	1.00	-0.59	-	-	-	0.43	1.00	-0.59	-	-	-
vanExel Model	0.49	0.75	-0.25	0.448	0.334	0.248	0.47	0.75	-0.20	0.510	0.317	0.238
Model (1)	0.41	0.66	-0.24	0.468	0.322	0.251	0.41	0.65	-0.19	0.476	0.305	0.253
Model (2)	0.41	0.64	-0.22	0.471	0.322	0.251	0.40	0.63	-0.22	0.475	0.316	0.251
Model (3)	0.41	0.78	-0.15	0.461	0.314	0.245	0.41	0.73	-0.15	0.550	0.316	0.245
Model (4)	0.42	0.76	-0.14	0.486	0.314	0.245	0.41	0.74	-0.14	0.550	0.317	0.245
Model (5)	0.41	0.64	-0.10	0.480	0.319	0.250	0.40	0.64	-0.10	0.544	0.311	0.244
Model (6)	0.42	0.70	-0.10	0.471	0.317	0.249	0.41	0.70	-0.10	0.549	0.313	0.241
Model (7)	0.42	0.80	-0.36	0.474	0.311	0.242	0.39	0.73	-0.38	0.555	0.316	0.251
Model (8)	0.42	0.78	-0.36	0.506	0.311	0.242	0.35	0.76	-0.40	0.554	0.326	0.268
Model (9)	0.41	0.64	-0.25	0.487	0.317	0.248	0.42	0.63	-0.25	0.551	0.309	0.243
Model (10*)	0.45	0.95	-0.52	0.382	0.350	0.244	0.47	0.94	-0.38	0.387	0.339	0.256
Model (11*)	0.45	1.00	-0.45	0.347	0.354	0.254	0.46	0.92	-0.32	0.365	0.343	0.261
Model (10#)	0.52	0.85	-0.42	0.409	0.369	0.281	0.50	0.81	-0.42	0.364	0.345	0.245
Model (11#)	0.51	0.69	-0.42	0.372	0.378	0.290	0.51	0.69	-0.42	0.360	0.341	0.249

*Denotes results obtained after Monte Carlo simulations

Denotes results obtained using the 'indirect method'

Table 9.11: Comparison between OLS and CLAD results - Within-Sample Prediction: Admission Data (Prediction dataset)

	Type of Regression Family	Mean (SD)	Min	P.25 [£]	Median	P.75 ^{&}	Max	% abs diff. = 0*	% abs diff. < 0.03\$	% abs diff. < 0.05#
Observed		0.42 (0.37)	-0.59	0.19	0.52	0.69	1.00			
Model (1)	OLS	0.42 (0.18)	-0.23	0.31	0.47	0.55	0.69	0	6.0	9.6
	CLAD	0.46 (0.22)	-0.28	0.32	0.52	0.62	0.75	0.7	7.5	12.3
Model (2)	OLS	0.42 (0.18)	-0.21	0.30	0.47	0.55	0.64	0	5.9	10.8
	CLAD	0.46 (0.21)	-0.31	0.33	0.53	0.63	0.73	0.5	6.3	13.1
Model (3)	OLS	0.42 (0.19)	-0.16	0.31	0.46	0.56	0.73	0	5.4	9.9
	CLAD	0.48 (0.21)	-0.21	0.33	0.53	0.63	0.79	0.9	8.3	14.0
Model (4)	OLS	0.42 (0.19)	-0.13	0.30	0.46	0.55	0.73	0	5.8	10.4
	CLAD	0.45 (0.22)	-0.34	0.31	0.53	0.63	0.80	1.7	8.6	13.6
Model (5)	OLS	0.42 (0.18)	-0.11	0.29	0.47	0.57	0.63	0	5.6	10.6
	CLAD	0.48 (0.17)	-0.16	0.39	0.54	0.64	0.71	2.9	9.4	13.9
Model (6)	OLS	0.42 (0.18)	-0.10	0.29	0.46	0.57	0.70	0	6.8	10.1
	CLAD	0.46 (0.23)	-0.11	0.30	0.54	0.64	0.78	3.9	8.5	12.6
Model (7)	OLS	0.42 (0.19)	-0.36	0.30	0.47	0.57	0.75	0	5.9	9.6
	CLAD	0.47 (0.23)	-0.45	0.46	0.55	0.63	1.00	1.9	10.3	15.6
Model (8)	OLS	0.42 (0.19)	-0.35	0.31	0.48	0.57	0.76	0	6.6	10.9
	CLAD	0.44 (0.23)	-0.52	0.35	0.55	0.66	0.89	3.0	8.9	14.4
Model (9)	OLS	0.42 (0.19)	-0.27	0.30	0.49	0.59	0.62	0	5.8	9.7
	CLAD	0.48 (0.23)	-0.42	0.35	0.56	0.69	0.71	3.6	8.0	12.9

* Percentage of Predicted EQ-5D scores whose values were identical to those of the observed EQ-5D scores

\$ Percentage of Predicted EQ-5D scores whose absolute difference from the observed EQ-5D score is less than 0.03

Percentage of Predicted EQ-5D scores whose absolute difference from the observed EQ-5D score is less than 0.05

£ 25% Percentile

& 75% Percentile

Table 9.12: Comparison between OLS and CLAD results - Within-Sample Prediction: Split Admission Data (Prediction dataset)

	Type of Regression Family	Mean (SD)	Min	P.25 [£]	Median	P.75 [*]	Max	% abs diff. = 0 [*]	% abs diff. < 0.03 ^{\$}	% abs diff. < 0.05 [#]
Observed		0.43 (0.36)	-0.59	0.19	0.52	0.69	1.00			
Model (1)	OLS	0.42 (0.18)	-0.24	0.32	0.46	0.55	0.69	0	6.7	10.8
	CLAD	0.51 (0.21)	-0.14	0.35	0.52	0.62	0.80	0.8	8.1	13.4
Model (2)	OLS	0.42 (0.17)	-0.19	0.30	0.47	0.55	0.64	0	6.6	12.1
	CLAD	0.49 (0.20)	-0.18	0.33	0.53	0.63	0.72	0.3	6.7	14.2
Model (3)	OLS	0.42 (0.18)	-0.15	0.31	0.47	0.56	0.70	0	5.1	10.6
	CLAD	0.47 (0.21)	-0.23	0.35	0.53	0.63	0.78	0.8	8.4	14.8
Model (4)	OLS	0.42 (0.18)	-0.08	0.31	0.47	0.56	0.70	0	6.0	11.2
	CLAD	0.47 (0.22)	-0.26	0.32	0.53	0.63	0.86	1.6	9.2	14.9
Model (5)	OLS	0.42 (0.17)	-0.08	0.31	0.46	0.56	0.62	0	5.9	11.5
	CLAD	0.48 (0.17)	-0.20	0.39	0.54	0.64	0.69	2.6	10.2	14.8
Model (6)	OLS	0.43 (0.17)	-0.06	0.30	0.46	0.57	0.68	0	6.7	10.7
	CLAD	0.46 (0.22)	-0.23	0.30	0.54	0.64	0.98	3.7	8.7	13.4
Model (7)	OLS	0.43 (0.18)	-0.32	0.32	0.47	0.56	0.75	0	6.1	10.4
	CLAD	0.49 (0.23)	-0.16	0.46	0.55	0.63	0.85	2.2	10.7	16.5
Model (8)	OLS	0.42 (0.18)	-0.31	0.32	0.48	0.56	0.75	0	7.3	11.9
	CLAD	0.49 (0.22)	-0.56	0.36	0.55	0.66	1.00	2.6	8.7	14.7
Model (9)	OLS	0.42 (0.18)	-0.24	0.32	0.48	0.57	0.62	0	6.4	10.2
	CLAD	0.48 (0.23)	-0.30	0.36	0.56	0.67	0.69	3.5	8.0	13.3

* Percentage of Predicted EQ-5D scores whose values were identical to those of the observed EQ-5D scores

\$ Percentage of Predicted EQ-5D scores whose absolute difference from the observed EQ-5D score is less than 0.03

Percentage of Predicted EQ-5D scores whose absolute difference from the observed EQ-5D score is less than 0.05

£ 25% Percentile & 75% Percentile

Table 9.13: Comparison between OLS and CLAD results - Within-Sample Prediction: Split Admission Data (Validation dataset)

	Type of Regression Family	Mean (SD)	Min	P.25 [£]	Median	P.75 ^{&}	Max	% abs diff. = 0*	% abs diff. < 0.03\$	% abs diff. < 0.05#
Observed		0.42 (0.38)	-0.59	0.12	0.55	0.69	1.00			
Model (1)	OLS	0.41 (0.20)	-0.19	0.28	0.45	0.56	0.65	0	4.2	6.3
	CLAD	0.48 (0.21)	-0.14	0.34	0.53	0.64	0.75	0.7	6.3	9.8
Model (2)	OLS	0.40 (0.19)	-0.22	0.25	0.46	0.55	0.63	0	4.2	7.7
	CLAD	0.46 (0.20)	-0.16	0.30	0.54	0.62	0.71	1.0	5.3	10.9
Model (3)	OLS	0.41 (0.20)	-0.15	0.27	0.44	0.56	0.73	0	6.7	8.4
	CLAD	0.46 (0.23)	-0.23	0.33	0.54	0.64	0.78	1.8	8.8	12.3
Model (4)	OLS	0.41 (0.20)	-0.14	0.27	0.44	0.55	0.74	0	5.6	8.8
	CLAD	0.46 (0.23)	-0.24	0.31	0.54	0.64	0.79	1.4	7.4	10.5
Model (5)	OLS	0.40 (0.20)	-0.10	0.26	0.45	0.57	0.64	0	5.3	8.4
	CLAD	0.46 (0.21)	-0.11	0.32	0.51	0.64	0.69	3.9	7.7	11.9
Model (6)	OLS	0.41 (0.20)	-0.10	0.28	0.45	0.57	0.70	0	4.9	8.8
	CLAD	0.45 (0.22)	-0.08	0.26	0.52	0.64	0.69	4.2	8.4	10.9
Model (7)	OLS	0.39 (0.20)	-0.38	0.26	0.43	0.54	0.73	0	5.6	7.7
	CLAD	0.44 (0.24)	-0.53	0.35	0.54	0.63	0.74	1.0	9.8	14.0
Model (8)	OLS	0.35 (0.19)	-0.40	0.24	0.37	0.49	0.76	0	5.3	8.4
	CLAD	0.40 (0.23)	-0.29	0.34	0.53	0.63	0.76	4.2	10.2	14.4
Model (9)	OLS	0.42 (0.20)	-0.25	0.24	0.45	0.59	0.63	0	4.6	9.1
	CLAD	0.48 (0.20)	-0.16	0.36	0.57	0.67	0.69	3.9	8.8	12.6

* Percentage of Predicted EQ-5D scores whose values were identical to those of the observed EQ-5D scores

\$ Percentage of Predicted EQ-5D scores whose absolute difference from the observed EQ-5D score is less than 0.03

Percentage of Predicted EQ-5D scores whose absolute difference from the observed EQ-5D score is less than 0.05

£ 25% Percentile & 75% Percentile

Table 9.14 Differences between OLS and CLAD Predicted EQ-5D utility scores

	Within-Sample Prediction: Admission Data (Prediction dataset)				Within-Sample Prediction: Split Admission Data (Prediction dataset)			
	Mean (SD) Predicted EQ-5D [OLS]	Statistical test of difference between Mean Predicted EQ-5D [OLS] and observed EQ5D*	Mean (SD) Predicted EQ-5D [CLAD]	Statistical test of difference between Mean Predicted EQ-5D [CLAD] and observed EQ5D*	Mean (SD) Predicted EQ-5D [OLS]	Statistical test of difference between Mean Predicted EQ-5D [OLS] and observed EQ5D*	Mean (SD) Predicted EQ-5D [CLAD]	Statistical test of difference between Mean Predicted EQ-5D [CLAD] and observed EQ5D*
Observed	0.42 (0.37)		0.42 (0.37)		0.43 (0.36)		0.43 (0.36)	
Model (1)	0.42 (0.18)	Z = -2.276 (p=0.023)	0.46 (0.22)	Z = -2.083 (p=0.037)	0.42 (0.18)	Z = -2.092 (p=0.036)	0.51 (0.21)	Z = -1.916 (p=0.055)
Model (2)	0.42 (0.18)	Z = -2.299 (p=0.021)	0.46 (0.21)	Z = -3.038 (p=0.002)	0.42 (0.17)	Z = -2.091 (p=0.036)	0.49 (0.20)	Z = -2.762 (p=0.006)
Model (3)	0.42 (0.19)	Z = -2.255 (p=0.024)	0.48 (0.21)	Z = -2.904 (p=0.003)	0.42 (0.18)	Z = -2.032 (p=0.042)	0.47 (0.21)	Z = -2.714 (p=0.007)
Model (4)	0.42 (0.19)	Z = -2.273 (p=0.023)	0.45 (0.22)	Z = -1.273 (p=0.203)	0.42 (0.18)	Z = -2.020 (p=0.043)	0.47 (0.22)	Z = -1.316 (p=0.188)
Model (5)	0.42 (0.18)	Z = -2.382 (p=0.017)	0.48 (0.17)	Z = -5.428 (p<0.001)	0.42 (0.17)	Z = -2.241 (p=0.025)	0.48 (0.18)	Z = -4.723 (p<0.001)
Model (6)	0.42 (0.18)	Z = -2.411 (p=0.016)	0.46 (0.23)	Z = -1.289 (p=0.1975)	0.43 (0.17)	Z = -2.313 (p=0.021)	0.46 (0.22)	Z = -1.200 (p=0.230)
Model (7)	0.42 (0.19)	Z = -2.33 (p=0.020)	0.47 (0.23)	Z = -4.268 (p<0.001)	0.43 (0.18)	Z = -2.074 (p=0.038)	0.49 (0.23)	Z = -3.797 (p<0.001)
Model (8)	0.42 (0.19)	Z = -2.353 (p=0.019)	0.44 (0.23)	Z = -2.930 (p=0.003)	0.42 (0.18)	Z = -2.073 (p=0.038)	0.49 (0.22)	Z = -2.778 (p=0.005)
Model (9)	0.42 (0.19)	Z = -2.437 (p=0.015)	0.48 (0.23)	Z = -3.361 (p<0.001)	0.41 (0.18)	Z = -2.230 (p=0.003)	0.48 (0.23)	Z = -2.992 (p=0.003)

*Wilcoxon signed rank sum test

Table 9.15: Comparing items in the Barthel ADL Index and the EQ-5D

Barthel	EQ-5D
<i>EQ-5D items directly comparable</i>	
Feeding	Self-care/Usual activities
Bathing	Self-care
Grooming	Self-care
Dressing	Self-care
Bowels	Self-care/Usual activities
Bladder	Self-care/Usual activities
Toilet Use	Self-care
Transfers	Mobility
Mobility	Mobility
Stairs	Mobility
<i>EQ-5D items not directly comparable</i>	
	Pain/discomfort
	Anxiety/depression

9.5 Comparison between CLAD and OLS models

As indicated in section 9.2.3, both the CLAD and OLS regression approaches were used to predict EQ-5D utility scores in models 1 to 9 so as to compare the performance of the two approaches. A summary of the key goodness-of-fit statistics obtained from the OLS models are shown in Tables 9.9 and 9.10. The results based on the discharge validation dataset are not discussed because of the unsuitability of this dataset as a validation dataset as discussed in sections 9.2.4 and 9.4.3.

9.5.1 Within-Sample Prediction: Admission Data (Prediction Dataset)

The results obtained from the ‘within-sample predictions’ based on the entire admissions data indicate that all OLS models had a predicted mean EQ-5D tariff score which was equal to the observed EQ-5D tariff score (0.42). These mean EQ-5D values therefore suggest that the OLS predicted scores were less biased (i.e. closer to the observed values) than those obtained from the CLAD regression models. While

the observed EQ-5D tariff score had a ‘full’ range (-0.59 to 1.00), none of the OLS models predicted EQ-5D values with such a wide range. The widest range for predicted EQ-5D scores amongst the models was for models (7) and (8) followed by models (1), (3) and (9) and the smallest range was again for model (5). These ranges were on average smaller than those for the CLAD model. This suggests that the CLAD models produced a better spread of predicted values than the OLS models. In terms of correlation, the results indicate that models (8), (3) and (7) predicted EQ-5D tariff scores that had the highest correlation with the observed EQ-5D tariff score (0.511, 0.500 and 0.499, respectively). On average, the OLS models predicted EQ-5D values that had a higher correlation to the observed EQ-5D values than those predicted by the CLAD model but the differences between the correlation values for the CLAD and OLS models for similar model specifications were at most 0.03. The RMSEs for the OLS models ranged from 0.311 to 0.325 while the MAEs varied from 0.243 to 0.251. Generally, the RMSEs for the OLS models were smaller than those for CLAD models while the MAEs for the former were found to be bigger than those for the latter. However, since the CLAD minimises the sum of absolute deviations, one needs to bear in mind that the MAE will tend to favour the CLAD model (Cheung et al. 2009). In order of performance measured by the combined RMSE and MAE scores, the results indicate that models (7), (8), (3), (4) and (9) had lower RMSE and MAE scores than other models. More weight was again placed on the RMSEs in determining performance (Hynman and Koehler, 2006).

9.5.2 Within-Sample Prediction: Split Admission Data (Prediction Dataset)

As shown in Table 9.10, the results obtained from the ‘within-sample predictions’ based on the split admissions data indicate that all OLS models had a predicted mean EQ-5D tariff score which was smaller or equal to the observed EQ-5D tariff score.

The OLS mean EQ-5D predicted values were within 3 percentage points of the observed EQ-5D values and were therefore less biased than those obtained from the CLAD regression models (which were within 5-21 percentage points of the observed EQ-5D values). While the observed EQ-5D tariff score had a ‘full’ range (-0.59 to 1.00), none of the OLS models produced predicted values with such a wide range. The widest range for predicted EQ-5D scores amongst the models was for models (7) and (8) followed by models (1), (3) and (9) and the smallest range was again for model (5). These ranges were on average smaller than those for the CLAD model. In terms of correlation, the results indicate that models (8), (3) and (7) predicted EQ-5D tariff scores that had the highest correlation with the observed EQ-5D tariff score (0.511, 0.500 and 0.499, respectively). Again on average, the OLS models predicted EQ-5D values that had a higher correlation to the observed EQ-5D values than those predicted by the CLAD model though these values from the two models did not differ by more than 0.06 for respective models. The RMSEs for the OLS models ranged from 0.311 to 0.325 while the MAEs varied from 0.243 to 0.251. Generally, the RMSEs for the OLS models were smaller than those for CLAD models while the MAEs for the former were found to be bigger than those for the latter. In order of performance measured by the combined RMSE and MAE scores, the results indicate that models (7), (8), (3), (4) and (9) had lower RMSE and MAE scores than other models.

9.5.3 Out-of-Sample Prediction: Split Admission Data (Validation Dataset)

The results obtained from the ‘out-of-sample predictions’ based on the split admissions data show that all except two OLS models (7 and 8) had a predicted mean EQ-5D tariff score which was bigger or equal to the observed EQ-5D tariff score (Table 9.10). The OLS mean EQ-5D predicted values ranged from 0 to within 9

percentage points of the observed EQ-5D values and were therefore appear to be less biased than those obtained from the CLAD regression models (range was between 5 and 14 percentage points of the observed EQ-5D values). While the observed EQ-5D tariff score had a ‘full’ range (-0.59 to 1.00), none of the OLS models produced predicted values with such a wide range. The widest range for predicted EQ-5D scores amongst the models was for model (9), followed by models (2), (6) and (8) and the smallest range were for models (1) and (5). These ranges were on average smaller than those for the CLAD model implying a limited spread of predicted EQ-5D values for the OLS model. In terms of correlation, the results indicate that models (7), (8) and (9) predicted EQ-5D tariff scores that had the highest correlation with the observed EQ-5D tariff score (0.326). On average, correlation between the EQ-5D utility values predicted by the OLS models and the observed EQ-5D values was similar to that between EQ-5D values predicted by the CLAD models and the observed EQ-5D values. The RMSEs for the OLS models ranged from 0.311 to 0.326 while the MAEs varied from 0.241 to 0.268. Five OLS models (2, 3, 4, 7 and 8) had higher RMSEs than their counterpart CLAD models while the MAEs for all OLS models were found to be bigger than those for CLAD models. In order of performance measured by the combined RMSE and MAE scores, the results indicate that models (7), (8), (3), (4) and (9) had lower RMSE and MAE scores than other models.

9.5.4 OLS or CLAD?

The results revealed mixed messages about the performance of the two models. Compared to the CLAD models, the OLS models on average predicted mean EQ-5D utility scores that were closer to the mean EQ-5D observed values, suggesting that these models produced mean EQ-5D values that were less biased. Indeed what is of

concern to an analyst when dealing with cost-effectiveness analysis is the group or aggregate mean EQ-5D scores rather than individual scores. It has actually been argued that, regardless of the degree of skewness, it is theoretically more correct to calculate mean utility based on the welfare economic principle that what should count is the strength of all individuals' preferences (Clarke et al. 2002). On the basis of this argument, the OLS model would therefore be preferable to the CLAD model.

Although the mean predicted EQ-5D scores based on the OLS models followed the mean observed EQ-5D scores more closely than those based on the CLAD models, the spread of the OLS predicted scores was limited. As shown in Table 9.11, the standard deviations (SDs) of the OLS scores (based on the entire admissions dataset) were on average only about 50% the size of those for the observed EQ-5D scores. On the other hand, the SDs for the CLAD models were on average about 60% the size of the SDs for the observed EQ-5D scores. This explains why the range for the OLS predicted scores was on average smaller than that of the CLAD and the observed EQ-5D scores. A similar pattern can be seen in the split admissions prediction dataset also shown in Table 9.12.

Examining the minimum, 25% percentile, median, 75% percentile and the maximum scores for the observed EQ-5D utility scores, also presented in Table 9.11, shows that the distribution of these measures of spread for the observed EQ-5D scores was more closely described or 'mimicked' by the CLAD predicted scores rather than by the OLS scores. This is more so for the latter three measures of spread (75% percentile, median and maximum). Tables 9.12 and 9.13 show that similar results were obtained from the split prediction and split validation datasets obtained after splitting the

admissions dataset into two as explained in section 9.2.4. These results were similar to those of other studies that compared the spread of the OLS and CLAD predicted EQ-5D utility scores (Cheung et al. 2009).

As presented in Table 9.14, the signed-rank test was also used to ascertain whether there were any statistically significant differences between the distribution of the observed EQ-5D scores on one hand and the OLS or the CLAD predicted values, on the other. The results showed that all of the OLS predicted scores were statistically different from the observed EQ-5D scores and this was the case for all datasets (complete admission dataset, split prediction dataset and split validation dataset). All of the differences between the observed EQ-5D scores and the CLAD predicted scores were also statistically significant with the exception of those in Models (4) and (6) for the full admission dataset and models (1), (4) and (6) for the split prediction dataset. Whilst both approaches performed badly on this score, the CLAD performed marginally better than especially for two of the models picked as better performing (models 4 and 6).

Another way of ascertaining bias in the predicted EQ-5D scores is to examine the residuals of the two models. The residuals were obtained by subtracting the predicted from the observed EQ-5D utility scores. The residuals were then plotted as histograms. Figures 9.8 to 9.15 show the plots of the OLS and CLAD residuals for the models that were deemed to have performed better than the others (models 3, 4, 6 and 7). The plots for the rest of the models are shown in Appendix A10. The plots appear not to show any significant differences between the residuals of the two models.

The analysis of bias was taken further and, as was done in Gray et al (2006), the percentage of predicted EQ-5D values that were equal to, or within a few units of, the observed EQ-5D scores was also determined. Definitions of what should be a MCID are debatable but as already indicated in section 9.4.3, values of 0.03, 0.05 and 0.081 have been suggested elsewhere. Therefore for the purposes of comparison, the two smallest values (0.03 and 0.05) were utilised. As shown again in Table 9.11, none of the OLS models predicted utility scores on their exact actual utility score. On the other hand, up to 3.9% of the utility values predicted by the CLAD model were equal to the actual observed utility scores. Further, only slightly over 5% and 9% of the OLS predicted values were on average within 0.03 and 0.05 of the observed values, respectively, compared to 8.4% and 13.6% for the CLAD. Similar results were obtained for the split admission prediction and validation datasets as shown in Tables 9.12 and 9.13. The CLAD model, on this aspect, therefore appeared to have performed better than the OLS models.

As revealed by the results in Tables 9.9 and 9.10, however, the OLS models were associated with lower RMSEs but higher MAEs compared to the CLAD model. This suggests that even though the CLAD models predicted a bigger proportion of scores that were closer to the observed EQ-5D values (i.e. improved model performance by increasing the number of correct estimates), the incorrect predictions were associated with a higher degree of error compared to the incorrect predictions from the OLS models. On this score therefore, the OLS model can be considered to have performed better.

As shown in Tables A5 and A6 in the appendix, the coefficients of determination (R-squared statistics) for the OLS models were higher than those for the CLAD models suggesting that the former had more explanatory power than the latter. However, it is important to take note that because the OLS approach minimises the sum of squared residuals, R-squared statistics will tend to favour the OLS rather than the CLAD models (Cheung et al, 2009).

The goal of predicting utilities is to be able to obtain group mean EQ-5D scores rather than individual scores, which can be used in cost utility analysis. Therefore a predictive model that gives group means that are close to the observed values would be ideal. However considering that utility scores are usually skewed in their distribution (skewness in the intermediate care dataset EQ-5D scores was confirmed by the Shaprio-Francia test - z-score = 15.1, $P < 0.001$), care should be taken to choose a model that predicts utility values which closely describe the variation and distribution of the utilities (Cheung et al. 2009). One should therefore consider how closely the predicted values match the other important measures of spread. On this criterion, the CLAD appears to have performed better than the OLS. In the face of heteroscedasticity, biased estimates would be obtained if the OLS were to be used (Greene, 1997; Gray et al. 2002; Payakachat et al. 2009). In the presence of heteroscedasticity, the CLAD model has been shown to be theoretically unbiased compared to the OLS (Greene, 1997; Clarke et al. 2002; Sullivan and Ghushchyan 2006). The Breusch-Pagan test ($\chi^2 = 60$, $p < 0.001$) confirmed the presence of heteroscedasticity in the demonstration dataset and therefore pointed to the CLAD as better model theoretically. .

Having compared the merits and demerits of both models, a decision was made to use the solution provided by the CLAD model. However, sometimes the attraction of a model that predicts mean EQ-5D scores that are closest to the observed mean EQ-5D values (less biased) may supersede the need to ensure that issues of variation in distribution, heteroscedasticity and skewness are accounted for. In such an instance, the OLS results would be more acceptable. If this were to be the case, then the OLS approach would even be more attractive than the CLAD because of a number of other practical advantages including the fact that the OLS method is supported by most statistical packages and the ease with which the joint significance of several regression coefficients in an OLS model can be tested e.g. using the F-test (Cheung et al. 2008). The full model results for the OLS are also presented in the Appendix A8 and A9.

Figure 9.8: Plot of Residuals - Model 3 (OLS), Admission dataset

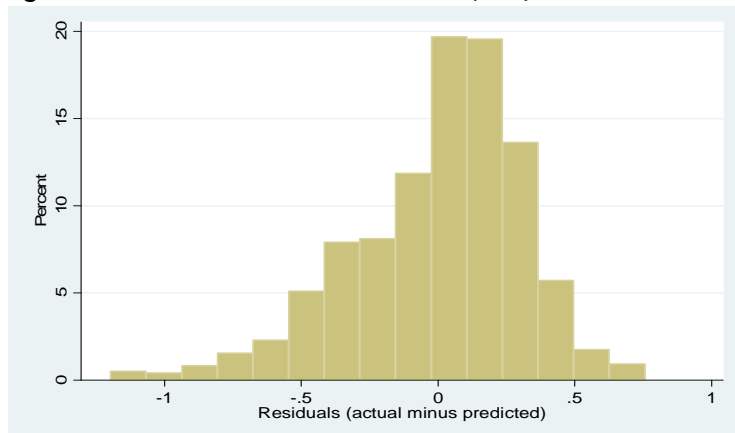


Figure 9.9: Plot of Residuals - Model 3 (CLAD), Admission dataset

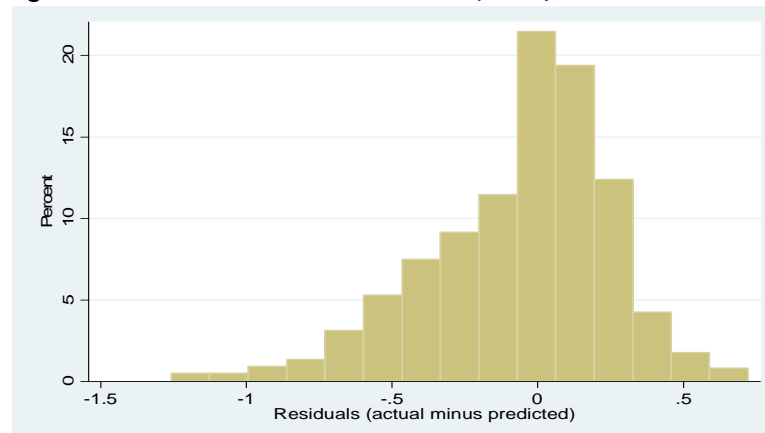


Figure 9.10: Plot of Residuals - Model 4 (OLS), Admission dataset

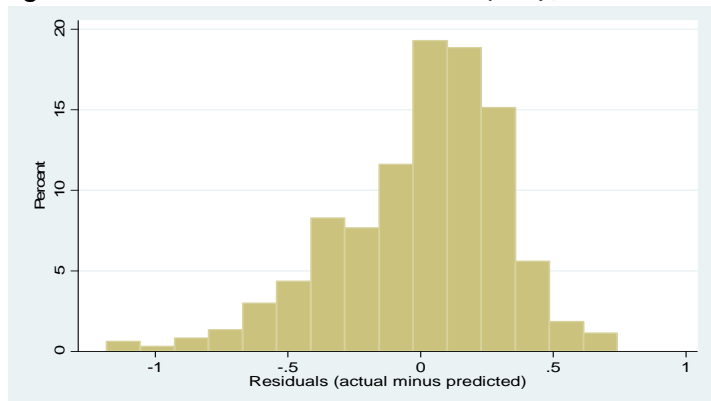


Figure 9.11: Plot of Residuals - Model 4 (CLAD), Admission dataset

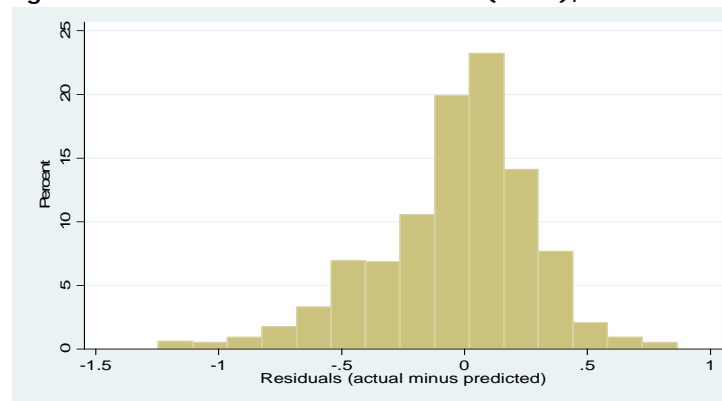


Figure 9.12: Plot of Residuals - Model 6 (OLS), Admission dataset

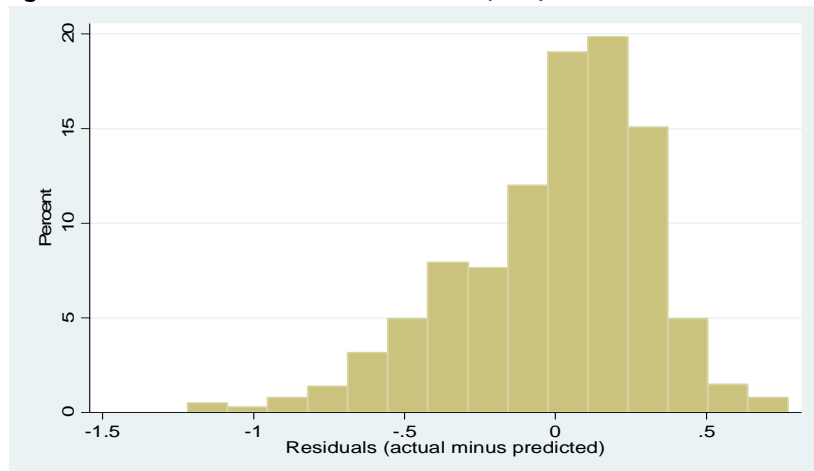


Figure 9.13: Plot of Residuals - Model 6 (CLAD), Admission dataset

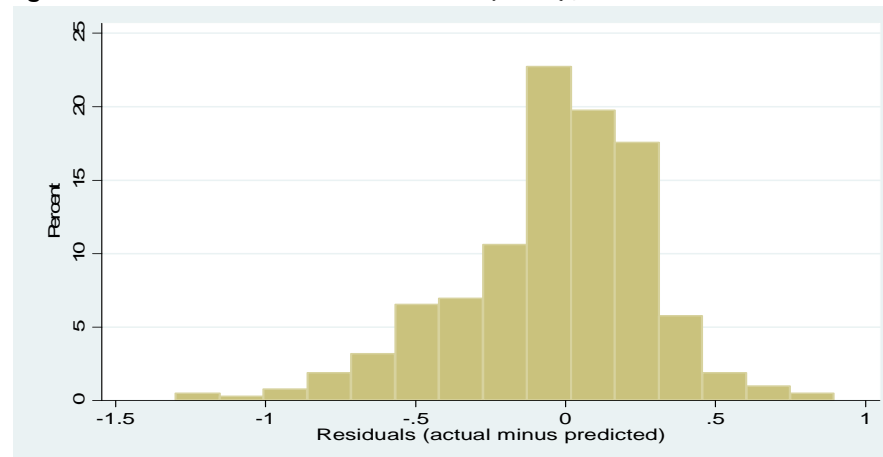


Figure 9.14: Plot of Residuals - Model 7 (OLS), Admission dataset

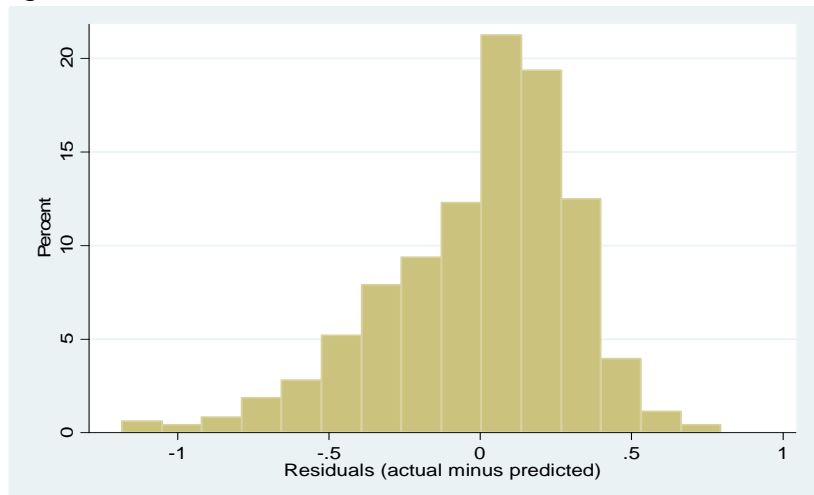
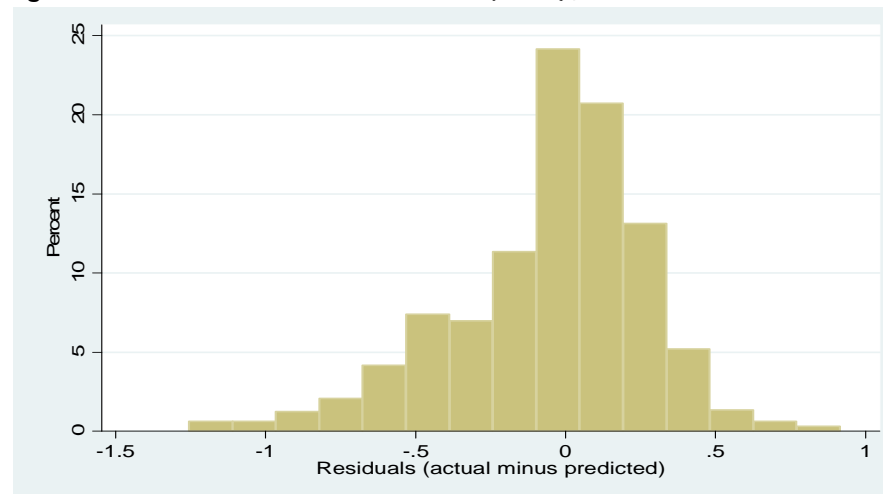


Figure 9.15: Plot of Residuals - Model 7 (CLAD), Admission dataset



9.6 Discussion

9.6.1 Key findings

Motivation for doing this analysis

As shown in chapter four, up to two-fifths of the information on the EQ-5D scores at admission (at the start of an intermediate care episode) was missing. The EQ-5D has been shown to be useful for generating utilities that can then be used in conducting cost effectiveness analyses (van Exel et al. 2004; Gold et al. 1996). A method of predicting EQ-5D scores from another outcome measure would therefore become valuable in cases where the EQ-5D scores are not available. . The other outcome measure that was collected in the national evaluation was the Barthel index. Importantly, there was less missingness in the Barthel scores compared to the EQ-5D i.e. only 31% of the Barthel scores were missing at admission. This therefore presented an opportunity to predict the EQ-5D scores from the Barthel scores.

What was shown in the analysis

The literature review conducted in chapter eight revealed different approaches for predicting utility-based outcome measures from non-utility-based ones. One of these, the van Exel et al (2004) model, produced results that were similar to models (1) and (2) which had comparable model specifications. The algebraic expression for the regression equation obtained from model (2) could be depicted as:

$$EQ-5D = -0.24 + 0.05 Barthel \quad (9.6)$$

The relative percentage difference in the coefficients between equations (9.6) and (9.1) (van Exel et al. 2004 model) was about 4% for both the constant and the coefficient for the Barthel. These results show that there was not much difference

between the two models. However, the RMSE and MAE for the van Exel model were higher than those for model (2). It was therefore clear that other specifications of the EQ-5D model may have greater predictive power than the van Exel et al (2004) model. This was however to be expected as this was in a sense an external or out-of-sample validation of the van Exel et al (2004) model. It was therefore not surprising that the model did not perform as well as other models. Concern could also be raised that the van Exel model and model (2) were limited because a totally independent individual was not allowed by these models to reach a value of 1.00 on the EQ-5D tariff score. But this could also be a reflection of the fact that there are dimensions of the EQ-5D that were not explicitly measured by the Barthel index, namely the pain/discomfort and anxiety/depression dimensions.

The results from this analysis seem to suggest that there is a positive relationship between the EQ-5D and the Barthel score, as one would expect. This was shown by the positive sign for the overall Barthel score coefficient in models (1) and (2) which depicted a relationship where an individual with functional independence (higher Barthel score) was associated with a high quality of life.

The results show that EQ-5D has a reasonable ceiling effect (7%) hence the use of CLAD models to account for this effect (Sullivan & Ghushchyan 2006a; Huang et al. 2008; Arabmazar & Schmidt, 1981). Most of the statistically significant coefficients in the models where the Barthel dimensions were treated as independent variables had the expected sign i.e. there was a positive relationship between them and the EQ-5D. This implied that the more functionally independent an individual was, the higher the quality of life that the individual was likely to have. Positive coefficients in the

models where the Barthel dimensions were entered as categorical variables were also as one would expect. For instance, the significant result for the ‘Toilet- Independent’ level of the Barthel index of 0.184 in Table 9.3 meant that individuals who were independent when it came to toilet functions were more likely to be associated with a higher quality of life compared to those who were dependent on others to carry out this function. A few coefficients however had a negative coefficient such as the ‘Bladder’ dimension in models (3) and (4) which is not what one would expect. It is difficult to pinpoint what the reason for this result was.

The results from this analysis seem to suggest that adding demographic characteristics such as age and gender did not add much in terms of the changes to the coefficients as could be seen in models (2), (4) and (8). This became apparent when the results of models (2), (4) and (8) were compared to those of models (1), (3) and (7), respectively. The only difference between the two sets of models was that age and gender were included as covariates in models (2), (4) and (8). The sizes of coefficients for comparable models were similar while the RMSE and MAE were also not that different. Entering the Barthel index or its dimensions as either continuous or categorical variables did not change the result.

Using only a select number of the 10 Barthel dimensions as covariates had a marked effect on the results. In particular, reducing the number of covariates resulted in previously insignificant variables turning out to be significant. This could be seen when model (4) was compared against model (5) or when model (8) was compared against model (9). The values of the RMSE and MAE were also different, with the ones for the reduced models being lower thereby implying better predictive models.

The R-Square values for the reduced models were however lower than those for the fuller models and this may be due to the differences in the number of covariates.

The addition of interaction terms in the models may have increased the explanatory power of the model but it did not result in lower RMSE and MAE values as the results of model (6) seem to suggest. This therefore appears to negate the value of adding interaction terms to a model specification.

Whenever outcome measures are designed to capture different aspects of functional performance or health, there will always be questions about what aspects or parts of one outcome measure are related to those of another. This was the case in this analysis as the EQ-5D is a generic quality of life measure while the Barthel is a measure of functional independence. With multinomial logistic regression, it can be possible to identify which dimensions of the EQ-5D are more closely related to the Barthel or its dimensions. Apriori, certain dimensions of the EQ-5D were expected to be associated with particular Barthel dimensions as shown in Table 9.8. The results of the analyses in this chapter show that there was more association between the Barthel index and all its dimensions on one hand and the EQ-5D dimensions of 'Mobility', 'Self-Care' and 'Usual activities' as opposed to the 'Pain/Discomfort' and 'Anxiety/Depression' dimensions. This was borne out by the larger R-Squared values obtained for the regressions involving the first three EQ-5D dimensions in models (10*), (10#), (11*) and (11#).

The results of the goodness-of-fit analysis seemed to suggest that in order of performance measured by the combined RMSE and MAE scores models (3), (7), (4)

and (6) performed better than the other models. with models (3) and (4) preferred on account of parsimony.

The multinomial logistic regression models performed the worst in terms of predicting the EQ-5D tariff scores. Models (10*), (10#), (11*) and (11#) had the highest RMSE and MAE values amongst all the models. This is true whichever method one uses to predict EQ-5D tariff scores from multinomial logistic regression models. These results were similar to those found by other studies that compared the multinomial logistic regression model to OLS regression (Tsuchiya et al. 2001; Gray et al. 2006). These models also had the lowest correlation between predicted and observed EQ-5D tariff scores. Even though the multinomial logistic regression models did not perform as well as the others, they still showed that, as would be expected, certain EQ-5D domains were more likely to be associated with particular Barthel dimensions than others. In particular, three EQ-5D domains ('Mobility', 'Self-Care' and 'Usual Activities') were more likely to be associated with some Barthel dimensions. One would expect 'Mobility' to be associated with the 'Transfer', 'Mobility' and 'Stairs' dimensions of the Barthel index and this was the case. The relationship though between EQ-5D mobility and Barthel dimension of 'Bowels' was not easy to explain. The Barthel dimensions that were associated with the 'Self-Care' domain of the EQ-5D ('Transfer', 'Dressing', 'Stairs' and 'Bathing') were again as you would expect. The same can be said about the Barthel domains that were associated with the 'Usual activities' domain of the EQ-5D ('Toilet', 'Dressing', 'Stairs', 'Bathing' and 'Bladder'). The base category for each of the EQ-5D dimensions was the 'No Problems' level. As a result, a negative sign on the coefficients was what one would assume i.e. people with some or extreme problems on any of the EQ-5D dimensions

of mobility, self-care, usual activities, pain/discomfort or anxiety/depression were more likely to be associated with a lower quality of life compared to those with no problems on each of the EQ-5D dimensions. What these results seemed to underline was the fact that the Barthel index is essentially a measure of functional outcome rather than one of general health hence the strong relationship between the Barthel index and its dimensions on one hand and the EQ-5D functional-related domains on the other.

Further, the method used to predict EQ-5D tariff scores from multinomial logistic regression models had an impact on how biased the predicted EQ-5D tariff scores were (i.e. how far away from the mean observed EQ-5D tariff score the mean predicted EQ-5D tariff scores were) and also on the size of the prediction error. The Monte Carlo simulation method produced scores that were closer to the observed EQ-5D tariff scores than those predicted using the ‘indirect method’ (or even those predicted by the CLAD models) but had higher prediction errors. This seemed to suggest that though the Monte Carlo method produced models that on average had the biggest errors, their ‘bias’, which is loosely defined as the distribution of the mean predicted EQ-5D tariff scores around the observed EQ-5D tariff scores, was low. Though the ‘indirect method’, on average, produced lower predicted errors than the Monte Carlo method, these were still higher than those obtained when using the CLAD models. There therefore does not seem to be any comparative advantage in using multinomial logistic regression (indirect or Monte Carlo method) to predict EQ-5D tariff scores. In addition, the multinomial logistic regression assumes independence of the irrelevant alternatives (Greene 1997). But this can not be case when the categories involved are those of EQ-5D dimensions. This therefore calls into

further question the suitability of multinomial logistic regression for predicting EQ-5D tariff scores.

Why does this analysis matter?

Both the EQ-5D and Barthel have been independently validated as reliable instruments for use with older people (Coast et al. 1998a; Brazier et al. 1996; Yohannes et al. 1998; Sainsbury et al. 2005). Some studies have also addressed the use of the Barthel as a proxy for HRQoL (van Exel et al. 2004). The results of the descriptive statistical analysis and those from the regression models affirm the association of low (high) Barthel scores with very low (high) EQ-5D which is an indicator of comparable sensitivity of the two measures. The results from these analyses have shown that there is a positive relationship between the two measures: lower mean scores on the Barthel correspond to lower mean scores on the EQ-5D. In addition, the regression results have revealed that the Barthel has the ability to correctly predict EQ-5D scores, albeit mildly. This finding is useful for circumstances where data on utility measures are missing which is not uncommon among populations of older people where self-reported measures are concerned (Brazier et al. 1996; Hilari et al. 2007; Muus et al. 2009). The result showing the mild predictive power of the models is in line with the findings of systematic reviews that looked at prediction of utilities from non-utility measures (Revicki & Kaplan 1993, Brazier et al. 2007, Mortimer & Segal 2008). The results from the 11 models demonstrated fairly reliable methods of mapping the Barthel index onto EQ-5D tariff scores. This would allow for the derivation of estimates of utility. It is also important to point out that these methods would not be able to accurately predict utility scores for individuals as they were not designed to do so but would be useful for predicting the

mean utility of a cohort. The results from these analyses seemed to suggest that a simple mapping of Barthel dimensions onto the overall EQ-5D tariff score using a regression model such as a CLAD estimator was preferable to one that mapped the Barthel dimensions onto EQ-5D dimensions using an approach such as multinomial logistic regression. The Barthel dimensions could be entered as either continuous or categorical variables. The CLAD model has the added advantage of predicting EQ-5D scores from the Barthel index while accounting for the ceiling effects, heteroscedasticity and skewness (Sullivan & Ghushchyan 2006a; Huang et al. 2008; Arabmazar & Schmidt, 1981). While a method such as that of using multinomial logistic regression may be able to establish the relationship between a non-utility based measure such as the Barthel and the dimensions of a utility-based measure such as EQ-5D, these results seemed to suggest that when one is interested in the mean EQ-5D tariff score, which is what is required when conducting cost-effectiveness analysis, then a simple rather than a complex model specification may do the work better. This however needs to be tested on further datasets before a conclusive deduction can be made.

Though the EQ-5D has been validated for use in populations of older people and therefore used as the mapping target of the Barthel index, the results from this analysis may again bring into scrutiny questions about the suitability of the EQ-5D for valuing health states of older people.

9.6.2 Key messages for health services researchers from this chapter

The most fundamental message from this chapter is that it is possible to reasonably predict the EQ-5D from the Barthel using a regression model framework. The regression results obtained in this chapter suggest that it is preferable to use a simple

mapping of Barthel dimensions onto the overall EQ-5D tariff score as opposed to mapping onto EQ-5D dimensions using multinomial logistic regression. There also does not seem to be any added value to including demographic characteristics in the predictive model. The mapping functions for the models that were chosen as better performing were found to be stable across validation datasets as the differences in the predicted EQ-5D values, RMSEs and MAEs were at most 0.02, 0.016 and 0.008, respectively. These differences were found to be below the minimally important difference (MID) cited in some literature. The EQ-5D maximum differences cited above were below the MID for the EQ-5D of 0.03 suggested by Dobrez et al (2007) while the maximum RMSEs and MAEs were below the values that have been deemed to be acceptable in some studies: 0.017 for the RMSEs as used in Marshall et al (2008), Tsuchiya et al (2002), and Gray et al (2006) and 0.01 for the MAEs as shown in Franks et al (2004), Gray et al (2006) and Barton et al (2008b). Both the OLS and CLAD regression models were run and their results compared. The OLS models predicted EQ-5D values that were less biased (i.e. mean EQ-5D utility values closest to the observed mean EQ-5D utility values), had a higher correlation to the observed EQ-5D utility values and also had a lower RMSE implying a greater predictive ability. The CLAD model on the other hand had predicted EQ-5D utility values whose distribution and variation more closely described that of the observed EQ-5D utility values as measured by the standard deviation, median, 75% percentile and maximum values. Further, the CLAD models predicted a bigger proportion of EQ-5D utility values that were within 0.03 and 0.05 units of the observed EQ-5D values compared to the OLS models. In addition, some CLAD models predicted EQ-5D utility values whose distribution was not statistically different from that of the observed EQ-5D values (measured by the signed-rank test) while all of the EQ-5D utility values

predicted by the OLS had a distributions that were statistically significantly different from those of the observed. Lastly, the CLAD model has been shown to be theoretically unbiased in the presence of heteroscedasticity compared to the OLS model (Greene, 1997; Clarke et al. 2002; Sullivan and Ghushchyan 2006). As heteroscedasticity was confirmed in the demonstration dataset by the Breusch-Pagan test ($\chi^2 = 60$, $p < 0.001$), the CLAD was considered to be a better model theoretically. Because of the above reasons, the decision was made to use of the results of the CLAD regression model instead of those from the OLS model.

9.7 Conclusion

As Tsuchiya et al (2001) also point out, these results seem to suggest that it will always be preferable to have preference or utility-based instruments included in the study design and any mapping of non-utility-based instruments to preference-based indices should be a second best. However in situations where it was not possible to collect utility-based measures for one reason or another, using these methods will greatly help in calculation of utility indices. In particular, models (3), (4), and (7) are contenders for predicting EQ-5D tariff scores from Barthel indices. Because models (3) and (7) need age and gender to be specified in the model, these results suggest that model (4) is a better model as it does not rely on having information on demographic characteristics available before running them. These analyses also appeared to suggest that there is not much to be gained from the use of models, such as the multinomial logistic regression model, that allow for the prediction of individual EQ-5D profiles before determining the EQ-5D tariff scores. The next chapter discusses the implications of the various reviews and analyses presented in this thesis so far.

CHAPTER TEN – DISCUSSION

10.1 Summary of Findings

This thesis set out to identify statistical issues that are commonly associated with evaluations of services for older people with a view to establishing, and demonstrating the use of, the best methods for dealing with them. This was done in two stages:

First, a comprehensive literature review of studies that have reported such evaluations was conducted. Restrictive criteria were used in identifying appropriate studies in that only studies conducted in a UK setting and on populations of older people aged 65 or over were considered. The results from this review which were reported in chapter three highlighted at least seven major categories of statistical issues that can be present in a dataset drawn from a population of older people. Another review based on more relaxed criteria was done and reported in chapter eight. This was conducted to unearth more approaches, not revealed in chapter three, for dealing with some of the statistical problems.

Second, the dataset drawn from the national evaluation of costs and outcomes of intermediate care services for older people in the UK was examined to see if it had any of the statistical problems identified in chapter three. Three statistical problems were identified and robust approaches for dealing with them were then used. The approaches used were those identified in the studies reported in chapters three and eight.

It should however be noted that this thesis does not make any claim that these statistical problems are only found in populations of older people or indeed that the nature of the problems in these populations is different to those of younger populations. This may or may not be the case but a definite position can only be taken if a similar analysis as the one performed in this thesis was conducted on populations of younger people. Because the focus of this thesis was on evaluations of older people, generalised conclusion to other populations cannot be made. The following are the key findings from this thesis.

10.2 Intermediate care and catering for the needs of older people

The UK, like many OECD countries, has been experiencing an increase in the population of older people in both relative and absolute terms. Another observation made in this thesis was that of an even bigger increase in the proportion of the ‘oldest old’ (those above 75 years of age). Both trends were projected to continue in the future (United Nations 2006). Older people are faced with needs that are different from those of younger members of the population. In the UK, eight themes, espoused as the eight standards of the NSF for older people (Department of Health 2001b) were pointed out as ways of catering for the unique needs of older people. Intermediate care was identified as one of these standards. There is no consensus on the correct definition of intermediate care reflecting the complexity of the area. Adding to this complication are the many models of intermediate care services that were identified in the UK including rapid response, hospital at home, residential rehabilitation, supported discharge and day rehabilitation (Department of Health, 2001a). Many evaluations of intermediate care services in the UK have been conducted and this thesis has shown that these evaluations took on many forms including experimental

studies (controlled and uncontrolled trials), non-experimental analytic studies and descriptive studies. One of these evaluations (ICNET 2005) provided the dataset on which the empirical analyses reported in this thesis were conducted.

10.3 Common statistical problems in evaluations of services for older people

A number of statistical problems or challenges were identified in the datasets used by most of the quantitative evaluation studies reviewed in chapter three. This search was restricted to studies based on UK populations of older people aged 65 years or over. Using this strict exclusion and inclusion criteria, studies that included anyone under the age of 65 years were excluded even when the mean age for the whole sample was above 65 years. There was thus potential for more problems to be unearthed if this review were to be extended to non-UK and/or age-specific populations. However, though this exercise can never be claimed to have been very comprehensive, it has revealed a fairly large number of statistical challenges that can be part of a checklist for researchers to consider before conducting any quantitative evaluation on a population of older people (Table 3.3). Another vital objective of this exercise was to explore methods that can be used to address these statistical challenges. These methods were adapted from two sources: (a) studies that were part of the literature review in chapter three or (b) studies reviewed in chapter eight based on relaxed and expanded search criteria. The approaches adopted from chapter eight were based on populations of individuals of all ages and were not restricted to UK settings. The fact that it was possible to adapt approaches for coping with statistical problems from studies based on younger and/or non-UK populations seems to suggest that some of these statistical problems are not unique to populations of older people in the UK or to older people in general. The thesis concentrated on three statistical challenges because

they were common in a number of the studies reviewed and also present in the demonstration dataset obtained from the national evaluation of costs and outcomes of intermediate care for older people in the UK. These were problems associated with the distributional characteristics of variables, missing data and the need to predict utility outcome measures from non-utility ones.

In conducting the empirical analyses reported in chapters five, seven and nine, this dissertation sought to answer three questions: Why was it necessary to look at the techniques used? What was revealed by the results of the analyses conducted? Why did it matter to carry out the analyses in the manner they were conducted in the first place? Key messages for health services researchers emanating from the analyses conducted were also spelt out.

Common distributional problems associated with variables used in a regression framework include skewness, heteroscedasticity and presence of too many zeros. Outcome measures that have bounds have the added challenge of dealing with ceiling and floor effects. What this thesis has shown is that a lot of studies that have evaluated services for older people routinely dealt with the problems associated with distributional characteristics of variables using theoretically-sound methods.

One of the objectives of the quantitative analysis in the national evaluation of costs and outcomes of intermediate care for older people in the UK was to understand the factors that cause variability in outcomes using a regression framework. An examination of the data revealed that the two outcomes of interest (change in the EQ-5D and change in the Barthel index) were skewed and heteroscedastic in their error

terms. Skewness leads to biased means and standard deviations while heteroscedasticity is associated with inefficient estimators and biased standard errors (Gujarati, 1995). In addition, the outcome measures were found to have a ceiling effect, which if not taken into account would lead to wrong predictions (Cheung et al. 2009; Payakachat et al. 2009). It was therefore important that robust methods of dealing with these problems were used in the regression models.

Some of these approaches identified in the literature review conducted in chapter three were adapted and used on the demonstration dataset. In chapter five, the GLM was used to model change in EQ-5D and change in Barthel scores. In these regression models, the objective was to identify the specific variables that explain variation in the two outcome measures. The GLM regression model was chosen because it is robust to skewness and heteroscedasticity (McCullagh & Nelder, 1989). The GLM results were then compared to those from an OLS model (that did not account for the skewness or heteroscedasticity) and marked differences were revealed in terms of which variables were significant predictors of outcomes. In the analysis reported in chapter five, more weight was placed on the ‘twin’ problems of skewness and heteroscedasticity than on those of the ceiling effect and heteroscedasticity because it was felt that the objective of this regression analysis was explanation of the model as opposed to estimation of a model for prediction. In addition to the GLM being theoretically superior to handle the twin problems of skewness and heteroscedasticity compared to the OLS, the diagnostic tests also showed that the model fit of the GLM model was good and that therefore the results from this model could be trusted. The fact that different results and conclusions concerning which variables were significant predictors of changes in outcome measures could have been reached if GLM was not

used shows that the method used to deal with skewed and heteroscedastic dependent variables matters.

By far the most prevalent statistical challenge was dealing with missing data due to reasons such as non response or attrition. While the best scenario would be to ensure that any study design did not lead to any missing data at all, many studies still find themselves faced with the problem. In the demonstration dataset, up to 42% of observations on some variables were missing. Methods for dealing with missing data vary according to the mechanism responsible for the missingness. Using the wrong method could potentially lead to biased and underpowered results (Roderick et al. 2001; Schafer, 1997). The majority of studies reported in chapter three simply ignored the missing values (complete case analysis – assuming that data are MCAR) without any regard to potential biases. Others also conducted complete case analysis but demonstrated that the approach did not bias their results. The literature review showed that very few studies tended to use principled and theory-based methods such as multiple imputation: only one study (Kaambwa et al. 2008) used multiple imputation which assumes that the data are MAR. No studies were found to have assumed that the data were MNAR.

While information on why some of the data were missing data was available, chapter seven explored what happens when such information is ignored and methods for dealing with missing data are chosen arbitrarily. A regression framework was used to understand variation in outcomes (change in EQ-5D and Barthel index) and costs per patient. The methods used, with the missing mechanism assumed in parenthesis, were:

regression using complete case analysis (MCAR), regression on multiply imputed data (MAR) and Heckman selection models (MNAR).

From the extra information gathered, it was established that the missing cost data was MCAR. Results obtained when a method assuming that the mechanism responsible for the missing cost data was MAR were not significantly different from those based on the MCAR assumption. These findings suggested that in many realistic applications, MAR methods were robust to departures from MAR in the dataset (Schafer et al. 2002; David et al. 1986). However, the use of an MNAR-based method in the costs per patient model yielded results that were so different to those obtained when either MCAR or MAR were assumed. Different conclusions would therefore be reached if the MNAR assumption was made for the missing cost data. All three missing data mechanisms were revealed to be potential reasons for the missing outcome data and the results from the Δ EQ-5D and Δ Barthel models showed that the choice of mechanism did not have any significant effect on the results.

The results of the analyses in chapter seven showed that there should not be any arbitrary selection of assumptions behind data missing mechanisms (Cohen & Cohen 1983; Orme & Reis 1991; Curran et al. 1998). The decision about the method to be used must consider both the reasons for missing data obtained during the data collection process and hypothesis testing (Curran et al. 1998). This is especially so when the MNAR assumption is used when data are actually MCAR. It however remains a fact that it is not easy to determine with certainty what missing data mechanism is responsible for missing observations in a dataset. Nevertheless, it is reassuring that as Schafer et al (2002) claim and as was shown in chapter seven,

departures from MAR in many realistic applications are not big enough to effectively invalidate the results of an MAR-based analysis. There however does not seem to be consensus about how big departures from MAR need to be before a method premised on MAR can be deemed to be inappropriate.

As long as one has to collect data from populations of older people, there will always be the potential of the outcome data being inaccurate or missing. This is especially so if such outcome data are self-reported. The availability of other credible outcome data may present an opportunity for such data to be used to predict the inaccurate or missing outcome data using regression methods. In the demonstration dataset, 40% of the information on the EQ-5D scores at admission (at the start of an intermediate care episode) was missing. Because of the desirability of the EQ-5D for generating utilities that can then be used in conducting cost effectiveness analyses (van Exel et al. 2004; Gold et al. 1996), a method of predicting ED-5D scores from other outcome measures is valuable. The Barthel index, a non-utility based outcome measure, was collected in the national evaluation and this outcome measure had fewer missing observations (31%). This situation therefore lent itself to a mapping exercise that would predict the EQ-5D utility scores (utility-based measures) from the Barthel scores. No evidence of any mapping between these two outcome measures was found from the review of literature from the UK on populations of older people. However the literature review reported in chapter eight yielded studies that have predicted utility scores from non-utility outcome measures. Most of these studies however showed only poor to moderate correlation between the two types of outcome measures.

Adapting some of the methods identified in chapter eight to the demonstration dataset showed that there was a positive relationship between the Barthel index and EQ-5D, which is an expected result. What this thesis has also shown is that one can choose to not add demographic characteristics to a regression predictive model as they did not improve the predictive models at all. Further, it was also evident that when the interest is in the predicting the mean EQ-5D tariff score as is the case when one is interested in conducting cost-effectiveness or cost-utility analysis, then a simple regression model (such as CLAD model) rather than a more complex one (like the multinomial logistic model) will work better. These results are borne out by those of other studies such as Tsuchiya et al (2001) and Gray et al (2006). On the other hand, the results of this thesis suggest that an even simpler model such as that used in van Exel et al (2004), where the overall Barthel score was the only explanatory variable, may be outperformed by other models where Barthel dimensions are used as explanatory variables instead. The CLAD regression model was chosen for predicting overall EQ-5D tariff scores because of the greater weight placed on constraining the predicted EQ-5D score between its lower and upper limits. The results of the CLAD model were compared to those obtained using an OLS model and were found to be preferable on account of them being more robust, theoretically, to skewness, heteroscedasticity and ceiling effects. Further, the CLAD predicted EQ-5D scores were shown to have more closely described the distribution of the observed EQ-5D scores (as measured by the standard deviation, median, 75% percentile, maximum EQ-5D value and signed-rank test) and also predicted a larger proportion of EQ-5D utility values that were within 0.03 and 0.05 of the observed EQ-5D utility scores. The OLS model however also performed better in terms of the bias of the predicted EQ-5D scores, correlations and RMSE.

The fact that it was possible to obtain reasonable mapping functions for predicting EQ-5D utility scores from Barthel index scores is an important finding that would be useful in situations where the former is missing but information on the latter is available.

10.4 Contribution to literature and key messages to health services researchers from this thesis

The review of population trends among older people in the UK and worldwide brought out the key message that the population of older people as well as their unique needs has been growing. Many models of care for older people exist including intermediate care. A literature review was conducted in chapter two to consider the results of quantitative evaluations of intermediate care that had been undertaken in the UK since 2000. This cut-off was chosen so that the review could build on the work of another study (Parker et al. 2000) which was systematic review focussing on evaluative research literature on the costs, quality and effectiveness of different locations of care for older patients. At the close of chapter two, the key results from the post-2000 literature review of evaluations of intermediate care were reported. Very few economic evaluations have been conducted with most of the studies being non-experimental analytic studies. The key message from these evaluations was that there was insufficient evidence concerning the cost-effectiveness or effectiveness of intermediate care. This is because the results from the evaluations were mixed. Intermediate care was associated with both positive and negative outcomes. Some intermediate care services were no better or worse than alternative services. These results are in broad agreement with the conclusions of Parker et al (2000) and similar other studies (e.g. Shepperd and Illife, 2004; Lambert and Arblaster, 2000; Young

2002) and therefore add to this body of evidence. Beech et al (2004) submit that there is need for more research to be conducted so as to ascertain whether or not intermediate care services can be seen as effective, suitable and efficient alternatives to acute care. The lack of research evidence has led some to conclude that there is not enough scientific evidence on the benefits of intermediate care (Melis et al. 2004), which Beech (2005) attributed in part to the difficulty around using scientific methods when evaluating intermediate care schemes.

One of the main objectives of this thesis was to establish a statistical framework for use when one is confronted with data from a population of older people. The first part of this exercise involved reviewing 56 studies that have evaluated services for older people in the UK and identifying the statistical challenges that were present in these studies. This is the first time such an exercise has been carried out on this kind of population in the UK. A key message obtained from the results reported in chapter three showed that there are at least seven broad groups of statistical challenges one has to be aware of when dealing with quantitative data from populations of older people and these are shown in Table 3.3. In order of prevalence, these are: missing data, lack of generalisability, problems associated with the distributional characteristics of variables, sample size and lack of power problems as well as the need for predicted outcome variables. Other statistical problems are lack of causality, different types of biases (participant, response and selection biases), problems due to heterogeneity and censoring. The results from this thesis will add to the body of evidence that will guide health services researchers on what statistical problems one has to be on the lookout for when working with data obtained from populations of older people.

Another objective of the thesis was to demonstrate how the challenges could be addressed. The empirical analyses reported in this thesis were based on a dataset obtained from the largest evaluation of intermediate care done and published in the UK to date (ICNET 2005). This relevant dataset provided an opportunity to demonstrate how statistical challenges could be dealt with. The results of these empirical analyses will therefore add to the body of evidence that demonstrates how these statistical issues can be handled in datasets based on populations of older people. These empirical analyses were specific to a population of older people in the UK and were therefore a unique piece of work that can be built upon and extended to cover other population groups in the UK as well as those from outside the UK. Three of the seven major statistical problems identified in the literature review and also present in the demonstration dataset were tackled: problems associated with the distributional characteristics of variables, bias due to missing data and the need to predict utility outcome measures from non-utility ones.

The use of the GLM to deal with the twin variable distributional problems of skewness and heteroscedasticity is not an original approach even for this population of older people. However, its use on this unique dataset where changes in EQ-5D and Barthel were modelled as functions of a number of explanatory variables adds to the body of evidence that laud the use of this model for dealing with these distributional issues. The use of GLM in dealing with skewness and heteroscedasticity has been demonstrated in other studies (Kaambwa et al. 2008; Killian et al. 2002; Manning & Mullahy 2001; Manning 1998; Mullahy 1998 and Cantoni & Ronchetti 2004, among others). When the results of the GLM were compared against those of the OLS, differences were seen. More variables were found to be statistically significant in

explaining variation in the outcome measures in the OLS compared to the GLM while some variables that were previously significant in the GLM model are no longer significant. The key message from this analysis is that when determining which specific factors significantly explain variation in an outcome variable, the regression method chosen, especially in the face of heteroscedasticity and skewness, is not without consequence. In the analysis reported in chapter five, the GLM was deemed to be a better model on theoretical and goodness-of-fit grounds.

The key message from the review of the theory on missing data and missing data techniques was that the methods for dealing with missing data are premised on the mechanism behind the missingness. The comparison of results of methods assuming that missing data were MCAR, MAR or MNAR highlighted the importance of correctly accounting for missing data. Different assumptions about the missing mechanism led to a different type of analysis and therefore different results. Considering the difficulty associated with identifying the correct mechanism behind missingness, the best solution would be to make attempts not to have any missing data in the first place. However, where missing data can not be avoided, the key message is that no arbitrary selection of assumptions behind data missing mechanisms should be made. Where possible, this decision should be guided by information on the reasons for the missing data which will complement hypothesis-testing-based detection methods. The results also showed that MAR-based methods for dealing with missing data are robust to departures from MAR in many instances. The use of sound statistical methods of analysing missing data was recommended. These results add to the body of similar evidence from other sources (Schafer 1997; Curan et al. 1998;

Fielding et al. 2008; Foster & Fang 2004; Fielding et al. 2006; Chavance 2004; Fayers et al. 1998).

A literature review which sought to identify approaches or techniques that have been used to predict utility-based outcome measures from non-utility based outcome measures in a regression framework was conducted and reported in chapter eight. This comprehensive review revealed a huge body of work that has been done in this area and will therefore add to the body of evidence from similar reviews such as Revicki and Kaplan (1993), Brazier et al (2007) and Mortimer and Segal (2008). The key message from this review is that various regression-based techniques for mapping non-utility-based outcome measures onto utility-based outcome measures exist. However, there are still methodological debates about the appropriate regression family to use in what circumstances as well as the acceptable cut-off points to use for goodness-of-fit statistics.

When dealing with data from populations of older people, it is not always possible to collect data on self-reported utility based outcome measures such as the EQ-5D. In such instances, it would be impossible to conduct economic evaluations such as cost effectiveness analyses. Using a CLAD model, chapter nine showed how the EQ-5D can be predicted from the Barthel index. The results of the CLAD were compared to those of the OLS and the merits and demerits of both were explored. A decision was made to use the CLAD rather than the OLS solution because predicted EQ-5D values from the former more closely described variation of the observed EQ-5D scores (as measured by the standard deviation, median, 75% percentile, maximum values and signed-rank test). Further, the CLAD models predicted a bigger proportion of EQ-5D

utility values that were within 0.03 and 0.05 units of the observed EQ-5D values than the OLS models. Lastly, the CLAD models were also deemed to be theoretically unbiased in the presence of heteroscedasticity compared to the OLS models. This is the first time that this mapping exercise involving these two outcome measures and also using a CLAD model has been carried out on a UK-based population of older people. The choice and use of the CLAD model adds to the body of evidence which considers this regression family to be appropriate for predicting health related quality of life measures such as the EQ-5D where ceiling effects, heteroscedasticity and skewness were issues (e.g. Saarni et al. 2006a; Sullivan et al. 2005; Saarni et al. 2006b; Sullivan & Ghushchyan 2006a; Sullivan et al. 2008; Clarke et al. 2002 and Sullivan & Ghushchyan 2006b).

The key message from the analysis done in chapter nine is that the results seemed to suggest that it is possible to reasonably predict the EQ-5D utility scores from the Barthel index. The results also showed that the best model is one where the EQ-5D tariff score is predicted from all Barthel dimensions entered as continuous independent variables. Though it will always be preferable to have preference or utility-based instruments included in the study design, the work in chapter nine showed that in situations where it is not possible to collect utility-based measures for one reason or another, fairly reasonable mapping of non-utility-based instruments to preference-based indices is possible. The poor to moderate association that was shown between the EQ-5D and the Barthel index has also been seen in other studies comparing other utility to non-utility outcome measures (van Exel et al. 2003; Ritvo et al. 2005; Kulkarni, 2006; Nichol et al. 2001; Bansback et al. 2007; Bosch et al. 1996; Barofsky et al. 2004 and Tsuchiya et al. 2002).

The key messages from this thesis for health services researchers who may be working with datasets drawn from populations of older people or similar groups are summarised in bullet format below:

- There at least seven major categories of statistical problems that can be found in datasets based on populations of older people (Table 3.3). The focus on three of the seven major (categories of) statistical issues in this thesis was guided by the presence of the issues in the demonstration dataset.
- Several methods for dealing with these statistical problems exist in the literature based on populations of both older and non-older people.
- When the goal of a regression model is to determine which specific variables significantly explain variation in an outcome variable, then the regression model chosen is of prime importance especially in the presence of statistical problems related to the distribution of a variable such as skewness and heteroscedasticity. The statistical problems related to the distributional characteristics of the variables must first be determined before a choice about the regression method to be used is made.
- Statistical techniques for dealing with missing data are premised on the mechanism responsible for the missingness. Techniques that assume that data are MAR are robust to situations where missing data are MAR.

- Various regression based techniques are available for predicting utility-based outcome measures from non-utility-based outcome measures and there are methodological considerations that need to be taken when they are used.
- It is possible to reasonably predict the EQ-5D utility scores from the Barthel index using a CLAD regression approach presented in chapter nine and the mapping functions are reasonably stable across validation datasets.
- In predicting EQ-5D scores from the Barthel, the OLS model performed better in terms of bias of the predicted EQ-5D utility scores, correlation between predicted EQ-5D and observed EQ-5D utility scores and the RMSEs of the predictive models. The CLAD model on the other hand predicted EQ-5D utility values whose distribution and variation more closely described that of the observed EQ-5D utility values (as measured by the standard deviation, median, 75% percentile, maximum values and signed-rank test). In addition, compared to the OLS models, the CLAD models predicted more EQ-5D utility values that were within 0.03 and 0.05 units of the observed EQ-5D values and were also theoretically unbiased in the presence of heteroscedasticity. The decision to use the CLAD solution was informed by the need to have a theoretically sound regression method that also predicted EQ-5D values whose distribution closely matched that of the observed EQ-5D values.

10.5 Strengths of the research

A major strength of this thesis is that the empirical analyses reported in chapters five, seven and nine were based on original quantitative data in the demonstration dataset

which represents the largest cohort of intermediate care episodes upon which data have been collected for the purposes of an evaluation. Data on patient characteristics, descriptors of intermediate care services and descriptors of intermediate care-related services on a total of 2,253 intermediate care patients were used.

This thesis also describes and summarises the key quantitative results of the national evaluation of costs and outcomes of older people in the UK that was completed in 2005. This evaluation focussed on costs and health outcomes and produced important results most notably in terms of the need for rigorous patient selection on admission to intermediate care.

This thesis presents the first comprehensive literature review using robust searching methods that was conducted in order to identify statistical issues that can be present in datasets obtained from populations of older people in the UK. To do this, all studies that reported evaluations of services for older people aged 65 and over in the UK were surveyed. The results of this review provide a broad resource for identifying statistical problems that can be present in studies that have evaluated services for older people. In addition, whereas other studies such as Parker et al (2000) focussed on review of evidence on effectiveness of intermediate care, this study sought to examine methodological issues with the view to contributing significantly to methodological debates on the best way to analyse quantitative data from populations of older people.

The analyses reported in this thesis utilised principled and robust statistical approaches with strong theoretical underpinnings in dealing with some of the statistical issues identified in the both the literature review and the demonstration

dataset e.g. GLM regression model, CLAD regression model, multiple imputation methods and Heckman sample selection methods. This is the first time that approaches of dealing with this number of statistical problems have been demonstrated on a unique dataset drawn from a population of older people in the UK. The results obtained will therefore be an important addition to the body of evidence on quantitative evaluations of services for older people particularly when dealing with statistical problems.

Another significant strength was the use of the CLAD model in predicting EQ-5D tariff scores from Barthel scores which is the first time this has been done. The models gauged to be better at predicting EQ-5D index scores from the Barthel score will therefore provide reference for future work involving mapping between the two outcome measures.

10.6 Weaknesses of the research

The data used in the national evaluation was obtained from five case study sites in the UK. Some of them were operating within a similar geographical area and county-wide context. As a result, the results from this thesis may not be very generalisable to other geographical areas in the rest of the country despite them being based on the largest evaluation of intermediate care done and published in the UK to date. These results would therefore need to be tested on other datasets so as to establish the generalisability of the findings. Also, the lower age limit used to define ‘older’ people in this thesis was 65 years but this arbitrary choice may not be commonly acceptable as other studies have used 55 years as the lower age limit. The majority of studies however used 65 as the cut-off age when defining older people. The databases

searched in chapters two, three and seven were not exhaustive and therefore more information could have been obtained by expanding the search to other databases. In addition, the sample searched in chapter three was restricted to studies conducted in UK settings and therefore many more statistical problems that one would face when dealing with quantitative data from an evaluative study of older people world-wide may have been missed.

Some data could have been estimated better had there been enough information. For instance, the cost variable was calculated using a top-down approach which method almost certainly lead to under-representation of the true variability in costs within and between intermediate care services. In addition, only two outcome measures were collected in the study i.e. the Barthel index and the EQ-5D. It would have been informative to conduct the analysis in this thesis in the context of other measures of outcome that have been used on populations of older people.

Because of constraints that come with analysing data and writing up a PhD thesis, not all methods assuming that data were either MCAR, MAR or MNAR were tested in chapter seven. The use of more methods would have added more weight to the evidence reported in this chapter. However, the approaches used represent each of the broader groups of methods and one should therefore have confidence in the results obtained and inferences drawn. Because of a particular focus on evaluations of older people in this thesis, it was also not possible to address the question of whether the nature of all statistical problems identified in chapter three were unique to populations of older people and could not be found in populations of younger individuals.

10.7 Recommendations for the future

These are divided into recommendations for research and recommendations for policy.

10.7.1 Research

- There should be research in identifying mechanisms of missingness to prevent the arbitrary selection of assumptions responsible for missing data.
- Attempts should be made to address some of the statistical issues that have not been tackled in this thesis e.g. unreliability/uncertainty of data, lack of generalisability, lack of causality and sample size/lack of power.
- The demonstration dataset collected data on only two outcome measures i.e. EQ-5D and the Barthel index. Future research should consider using other measures of health-related quality of life and functional independence available which have been validated for use in populations of older people e.g. health utility index (HUI), short-form 36 (SF-36), Nottingham extended activities of daily living (ADL) scale, Functional Status Index (FIS) and Quality of well-being (QWB) index.
- There is need for more economic evaluations of services catering for older people to be conducted so that evidence on costs-effectiveness can be obtained.

- A combined split sample/discharge sample should be used to overcome the problems of non-independence of the current discharge dataset when compared to the admission dataset.

10.7.2 Policy

- If the government's policy is to target intermediate care resources towards patients and/or services associated with the largest gains in quality of life and/or improvements in functional status, then the results from this thesis suggest this would be met by focussing on those patients with the greatest need (lower baseline EQ-5D and Barthel scores).
- Compared to supported discharge intermediate care services, admission avoidance services were associated with greater gains in health outcomes (both functional and more general quality of life). There is thus strong support for a policy that would focus more resources on admission avoidance as opposed to supported discharge services.

10.8 Conclusions

This thesis has shown that many studies reporting quantitative evaluations of services for older people have statistical problems that need to be addressed if the results obtained are to be credible. If these statistical problems are not dealt with correctly, they may distort the findings of an evaluation. This thesis identified a number of statistical problems including those related to distributional characteristics of variables, missing data and the need to predict utility measures of outcome from non-utility ones. The thesis has shown that it is important to use methods with a sound

theoretical background to deal with the problems and has demonstrated the application of appropriate methods to deal with the identified statistical problems using data from a large national study evaluating intermediate care services for older people in the UK.

APPENDIX

A1 - Key terms used to search for literature in chapter two.

These were divided into five groups (please note that ‘*’ represents a truncation facility to account for variations of the search term):

1. *Some terms and conditions associated with older people*

- Activities of daily living
- ADL
- Age*
- Alzheimer
- Dementia
- Depend*
- Elder*
- Fall*
- Functional
- Geriatric*
- Independen*
- Old*
- Over 65
- Over 70
- Over 75
- Over 80
- Over 85
- Stroke*

2. *Some terms associated with Intermediate Settings (Used the Social care online to get words/phrases that are common to Intermediate care)*

- Admission avoidance
- Care
- Closer to home
- Day hospital
- HAH
- HaH
- Hospital at home
- Hospital from home
- Hostel care
- Hotel care
- Intermediate care
- National service framework
- Nurse-led*
- Nursing*
- Rapid response*
- Rehab*
- Residential
- Social*
- Supported discharge
- Swing beds
- Home care service

3. *Some terms associated with Evaluation*

- Analys*
- Apprais*
- Assess*
- Eval*
- Exam*
- Find*
- Investi*
- Research*

- Review*
- Stud*
- Survey*
- Valu*

4. *Some terms associated with UK/Britain*

- Britain
- British
- English
- England
- Ireland
- Irish
- Northern Ireland
- Scot*
- UK
- United kingdom
- Wales
- Welsh

5. *Some terms associated with Quantitative Evaluation*

- Audit*
- Economic*
- Fig*
- Quantitative*
- Result
- Statistic*

A2 - Key terms used to search for literature in chapter three.

All the search terms were used in chapter two (A1) were used here as well with the exception that group 2 was expanded. The following terms were added to group 2:

- Activities for elderly people
- Activities for older people
- Age concern
- Community*
- Home meals
- Integrated support
- Local authority
- Long term
- Mental health
- Seniors forum
- Sheltered housing
- Social work
- Support service
- Vulnerable

A3- Barthel ADL index questionnaire

The following are the guidelines given by the Internet stroke center (www.strokecenter.org):

1. The index should be used as a record of what a patient does, not as a record of what a patient could do.
2. The main aim is to establish degree of independence from any help, physical or verbal, however minor and for whatever reason.
3. The need for supervision renders the patient not independent.
4. A patient's performance should be established using the best available evidence. Asking the patient, friends/relatives and nurses are the usual sources, but direct observation and common sense are also important. However direct testing is not needed.
5. Usually the patient's performance over the preceding 24-48 hours is important, but occasionally longer periods will be relevant.
6. Middle categories imply that the patient supplies over 50 per cent of the effort.
7. Use of aids to be independent is allowed.

Scores for each dimension are added up to make a single Barthel index score between 0 and 100. This is sometimes standardised so that the score is between 0 and 20. The dimensions, levels and scores in the Barthel questionnaire are:

<u>Activity</u>	<u>Score</u>
-----------------	--------------

FEEDING

0 = unable

5 = needs help cutting, spreading butter, etc., or requires modified diet

10 = independent _____

BATHING

0 = dependent

5 = independent (or in shower) _____

GROOMING

0 = needs to help with personal care

5 = independent face/hair/teeth/shaving (implements provided) _____

DRESSING

0 = dependent

5 = needs help but can do about half unaided

10 = independent (including buttons, zips, laces, etc.) _____

BOWELS

0 = incontinent (or needs to be given enemas)

5 = occasional accident

10 = continent

BLADDER

0 = incontinent, or catheterized and unable to manage alone

5 = occasional accident

10 = continent

TOILET USE

0 = dependent

5 = needs some help, but can do something alone

10 = independent (on and off, dressing, wiping)

TRANSFERS (BED TO CHAIR AND BACK)

0 = unable, no sitting balance

5 = major help (one or two people, physical), can sit

10 = minor help (verbal or physical)

15 = independent

MOBILITY (ON LEVEL SURFACES)

0 = immobile or < 50 yards

5 = wheelchair independent, including corners, > 50 yards

10 = walks with help of one person (verbal or physical) > 50 yards

15 = independent (but may use any aid; for example, stick) > 50 yards

STAIRS

0 = unable

5 = needs help (verbal, physical, carrying aid)

10 = independent

TOTAL (0–100):

A4 - EuroQol EQ-5D (EQ-5D) questionnaire

The EQ-5D is classified into five dimensions which are in turn divided into three levels. The different levels from each dimension can be combined to define 243 health states. Using “sets of values” derived from general population samples, values on the different levels may be converted to a single score. The dimensions and levels are:

MOBILITY

I have no problems in walking about

I have some problems in walking about

I am confined to bed

SELF-CARE

I have no problems with self-care

I have some problems washing or dressing myself

I am unable to wash or dress myself

USUAL ACTIVITIES (e.g. work, study, housework family or leisure activities)

I have no problems with performing my usual activities

I have some problems with performing my usual activities

I am unable to perform my usual activities

PAIN / DISCOMFORT

I have no pain or discomfort

I have moderate pain or discomfort

I have extreme pain or discomfort

ANXIETY / DEPRESSION

I am not anxious or depressed

I am moderately anxious or depressed

I am extremely anxious or depressed

A5 – Results from the Heckman selection model

Table A1: Heckman Selection Models – Costs, EQ5D and Barthel Models

<i>Variables</i>		Cost Model (n = 717, 125 obs censored)		EQ-5D model (n = 1105, 417 obs censored)		Barthel Model (n = 1105, 392 obs censored)	
Episode	Age in 2003	-0.011	0.007				
Characteristics	Gender	0.038	0.149	-0.171	0.098	-0.084	0.097
	Lives alone	-0.057	0.140	-0.040	0.091	-0.069	0.090
	Barthel score at admission	0.065	0.019**	-0.014	0.013	-0.024	0.013
	EQ5D score at admission	-0.331	0.214	-0.248	0.144	-0.025	0.144
Descriptors of IC Service	Type of IC	-0.770	0.160**				
	Transferred before end of IC episode	0.090	0.515	5.861	0.277**	1.436	0.511**
	Completed IC episode	0.163	0.496	7.219	0.278**	2.539	0.506**
	Other IC Outcome	-0.077	0.587	6.608	0.312**	1.715	0.531**
	Patient Died (Reference. Group)						
	Duration of Stay	0.009	0.003**				
Descriptors of IC-related Services	Referral – Primary	-0.389	0.314	-0.051	0.224	-0.336	0.237
	Referral – Hospital	-1.309	0.319**	-0.579	0.210**	-1.012	0.225**
	Referral – Other	-1.001	0.444	-0.184	0.329	-0.256	0.354
	Referral – Social Workers (Reference Group)						
	Alternative to IC – Other	0.913	0.225**				
	Alternative to IC – Home	-0.071	0.177**				
	Alternative to IC – Hospital (Reference group)						
	Supported Discharge Service	1.468	0.196**	1.186	0.101**	1.158	0.099**

Interactions	Barthel score at admission*Type of IC	-0.132	0.034**				
	Transfer before IC end*Type of IC	-0.471	0.245	0.216	0.358	0.177	0.319
	Completed Episode*Type of IC			0.569	0.235	0.036	0.221
	Other IC Outcome*Type of IC	0.112	0.384	-0.232	0.856	-0.353	0.827
	Patient died*Type of IC (Reference group)	0.739	0.373*				
		-0.319	1.046				
	Supported Discharge Service*Type of IC			-0.202	0.263	0.223	0.240
	Mills Ratio/Lambda (w/o Interactions)	-3.402	1.242**	-0.284	0.046**	-1.662	0.652*
	Mills Ratio/Lambda (with Interactions)	-4.506	1.767**	-0.143	0.047**	-0.101	2.260*

* 5 % level of significance, ** 1% level of significance; IC = intermediate care

Dep. variable for cost model: cost per patient present (1 if yes, 0 otherwise)

Dep. variable for EQ-5D model: EQ5D score present (1 if yes, 0 otherwise)

Dep. variable for Barthel model: Barthel score present (1 if yes, 0 otherwise)

A6 – Search terms for literature review in chapter eight

The search terms were divided into three groups:

1. Search terms associated with regression prediction or mapping

- Regress
- Validity
- Compare
- Comparison
- Predict
- Forecast
- Calculate
- Cross-walk
- Reliability
- Assess
- Agree
- Valuation
- Map
- Proxy
- Proxies
- Converge
- Associat*

2. General search terms associated with utility and non-utility measures of outcome

- Outcome
- Effect
- Utility
- Psychometric
- Index
- Cardinal values
- Health value
- Clinimetric
- Cardinal
- Preference
- Profile
- Health status
- Functional*

3. Search terms for specific measures of outcomes

- EuroQol .
- EQ-5D
- HUI*
- Short Form*
- SF*
- Health Utilities Index
- SIP*
- Disability Index
- ECOS*
- Quality of well being
- QWB
- General Well-Being

- GWB
- Visual analogue scales
- VAS
- Standard gamble
- SG
- Time tradeoff

A7 – CLAD Regression Results obtained using the split Admissions Dataset

Testing the van Exel (2004) model

The van Exel (2004) model was tested on the prediction dataset to see how well it performed. Using this model, a mean predicted EQ-5D tariff score of 0.50 (minimum of -0.25 and maximum of 0.75) was obtained compared to the 0.43 that was obtained for the observed overall EQ-5D tariff score. The RMSE score for the model was 0.334 while the MAE was 0.248. The results affirmed the association of low (high) Barthel scores with very low (high) EQ-5D tariff scores which is an indicator of comparable sensitivity of the two measures.

Appendix Tables A2, A3 and A4 show the results of the 11 regression models described in section 9.4.2 of chapter nine and performed on two-thirds of the split Admissions dataset.

Model 1

Table A2 shows a positive relationship between the overall EQ-5D tariff score and the Barthel overall score (a positive sign for the overall Barthel score coefficient). Age and gender were not significantly related to the overall EQ-5D tariff score. The coefficient of determination (R-squared) was 0.138.

Model 2

This model was the same as model (1) with the exception that age and gender were removed as independent variables. Here again as shown in Table A2, the overall

Barthel score was positively related to the overall EQ-5D tariff score and the R-squared was 0.134.

Model 3

As shown in Table A2, only two Barthel dimensions were statistically significant in this model. These were 'Dressing' and 'Stairs' and both were positively related to the overall EQ-5D tariff score. Age and gender were again not statistically significant in this model. The R-squared statistic was again higher than for the first two models at 0.166.

Model 4

When age and gender were excluded from model (3), the covariates that were statistically significant in the model are again significant in this model. In addition, however, 'Bladder' is also now significant. 'Dressing' and 'Stairs' also had the same positive signs as before while 'Bladder' was negatively related to the overall EQ-5D score. The sizes of the coefficients were similar but the coefficient of determination was slightly lower at 0.163.

Model 5

In this model, stepwise regression was used to reduce the number of covariates used in model (4). This resulted in only four variables being included in the model: 'Transfer', 'Mobility', 'Dressing' and 'Stairs'. All four covariates were statistically significant and also positively related to the overall EQ-5D tariff score as indicated in Table A2. The coefficient of determination was 0.155.

Model 6

Interaction terms were added to model (5) and the result was model (6). None of the main effects and interaction terms were statistically significant but the R-square was 0.163.

Model 7

In Table A3, four levels of Barthel dimensions are shown to have been statistically significant in explaining the overall EQ-5D tariff score. These were the 'Needs help' and 'Independent' for the 'Dressing' dimension and 'Needs help' and 'Independent' levels for the 'Stairs' dimension. The reference category for all of the categorical explanatory variables was the 'dependent/unable' category. All of the statistically significant variables had a positive relationship with EQ-5D. The R-square was 0.175.

Model 8

When age and gender were removed from the preceding model, all of the variables that were significant in that model are still significant. The R-squared is slightly lower though at 0.171.

Model 9

This was a reduced version of Model (8) arrived at using stepwise regression. The results showed in Table A3 show that all the levels that were significant in model (8) are again significant in this model. However, the 'Wheelchair independent', 'Needs help' and 'Independent' levels for the 'Mobility' dimension which were previously not statistically significant are now statistically significant. An R-squared value of 0.160 was obtained.

Model 10

Table A4 shows the results of the multinomial logistic regression results. While the coefficients in this model are not easy to interpret (Greene 1997), what they show is that some EQ-5D dimensions are more likely to be associated with particular Barthel dimensions. The EQ-5D 'Mobility' dimension is likely to be associated with the 'Mobility' and 'Stairs' dimensions. The Barthel dimensions that were associated with the 'Self-Care' domain of the EQ-5D were 'Grooming' , 'Dressing', 'Stairs' and 'Bathing'. These again are as one would expect. As expected also, the 'Toilet', 'Dressing', 'Stairs' and 'Bathing' dimensions of the Barthel index were associated with the 'Usual Activities' domain of the EQ-5D. The last two domains of the EQ-5D (Pain/Discomfort and Anxiety/Depression) do not have obvious counterparts among the Barthel dimensions (Table 9.8). Only one Barthel dimension (Bladder) was significantly associated with the former EQ-5D domain while only the 'Stairs' Barthel dimension was statistically significant in explaining the latter EQ-5D domain. Nearly all of the statistically significant coefficients had negative signs. The only exceptions were the 'Toilet' dimension in the 'Usual activities' multinomial regression model and the 'Bladder' Barthel dimension in the Pain/Discomfort model. The R-squared values with associated models in parentheses were 0.141 (Mobility), 0.197 (Self-Care), 0.120 (Usual Activities), 0.048 (Pain/Discomfort) and 0.059 (Anxiety/Depression).

Model 11

The results also presented in Table A4 show that the overall Barthel score is statistically related to all EQ-5D dimensions. All coefficients are negative and the R-squared values with associated models in parentheses were 0.010 (Mobility), 0.149

(Self-Care), 0.080 (Usual Activities), 0.029 (Pain/Discomfort) and 0.042 (Anxiety/Depression).

Table A2: CLAD model – Barthel dimensions entered as continuous variables: Split Admission Data

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Variables	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Total Barthel score	0.046*	0.004	0.046*	0.004								
Age	0.028	0.013			0.020	0.013						
Gender (1 = female, 0 = male)	-0.041	0.026			-0.039	0.025						
Barthel – Grooming					-0.040	0.047	-0.033	0.044				
Barthel – Toilet					0.105	0.050	0.103*	0.046				
Barthel – Feeding					0.059	0.036	0.061	0.046				
Barthel – Transfer					0.073	0.037	0.083	0.051	0.090*	0.044	0.026	0.081
Barthel – Mobility					0.038	0.032	0.036	0.033	0.071*	0.028	0.051	0.069
Barthel – Dressing					0.105*	0.031	0.104*	0.030	0.105*	0.019	0.110	0.087
Barthel - Stairs					0.046*	0.012	0.052*	0.011	0.052*	0.011	0.150	0.106
Barthel – Bathing					0.020	0.027	-0.000	0.024				
Barthel – Bladder					-0.042	0.025	-0.060*	0.026				
Barthel – Bowels					-0.003	0.041	-0.016	0.042				

Barthel - Transfer x Toilet											0.039	0.050
Barthel - Mobility x Toilet											0.006	0.049
Barthel - Dressing x Toilet											-0.011	0.044
Barthel - Stairs x Toilet											-0.051	0.053
Constant	-0.253*	0.084	-0.206*	0.068	-0.180	0.101	-0.119	0.101	-0.108	0.076	-0.082	0.098
R-Squared		0.138		0.134		0.166		0.163		0.155		0.163

Dependent Variable; EQ-5D * Statistically significant at 5% level

Table A3: CLAD model – Barthel dimensions entered as categorical variables: Split Admission Data

	Model 7		Model 8		Model 9	
Variables	Coeff.	SE	Coeff.	SE	Coeff.	SE
Age	0.022	0.014				
Gender (1 = female, 0 = male)	-0.045	0.025				
Grooming - independent	-0.033	0.035	-0.029	0.036		
Toilet – needs help	0.078	0.010	0.039	0.108		
Toilet – independent	0.157	0.128	0.111	0.127		
Toilet – dependent (Reference category)						
Feeding – needs help	-0.230	0.232	-0.230	0.211		
Feeding - independent	-0.172	0.232	-0.131	0.212		
Feeding – unable (Reference category)						
Transfer – major help	0.275	0.212	0.081	0.197	0.161	0.162
Transfer – minor help	0.326	0.202	0.164	0.194	0.186	0.145
Transfer – independent	0.406	0.198	0.263	0.197	0.273	0.155
Transfer – unable (Reference category)						
Mobility – wheelchair independent	0.009	0.182	0.018	0.205	0.002*	0.196
Mobility – needs help	0.158	0.106	0.122	0.092	0.164*	0.084
Mobility – independent	0.151	0.121	0.121	0.101	0.198*	0.077
Mobility – immobile (Reference category)						
Dressing – needs help	0.206*	0.066	0.183*	0.067	0.192*	0.067
Dressing – independent	0.298*	0.067	0.281*	0.073	0.279*	0.063
Dressing – dependent (Reference category)						
Stairs – needs help	0.086*	0.035	0.084*	0.037	0.088*	0.035
Stairs – independent	0.087*	0.027	0.101*	0.024	0.104*	0.021
Stairs – unable (Reference category)						

Bathing – independent	0.043	0.030	0.005	0.024		
Bladder – occasional accident	-0.054	0.064	-0.081	0.071		
Bladder – continent	-0.097	0.052	-0.126	0.069		
Bladder – incontinent (Reference category)						
Bowels – occasional accident	-0.033	0.118	0.022	0.113		
Bowels – continent	0.010	0.114	0.025	0.112		
Bowels – incontinent (Reference category)						
Constant	-0.182	0.248	0.070	0.234	- 0.164*	0.136
R-Squared		0.175		0.171		0.160

Dependent Variable; EQ-5D * Statistically significant at 5% level

Table A4: Multinomial Logit using Barthel dimensions as independent variables: Split Admission Data (Models 10 & 11)

	Mobility				Self-Care				Usual Activities				Pain/Discomfort				Anxiety/Depression			
	Some Problems		Extreme Problems		Some Problems		Extreme Problems		Some Problems		Extreme Problems		Some Problems		Extreme Problems		Some Problems		Extreme Problems	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Grooming	0.173	0.301	0.599	0.487	-0.556*	0.255	-0.258	0.380	-0.508	0.366	-0.067	0.388	-0.129	0.235	-0.121	0.318	-0.395	0.219	-0.284	0.331
Toilet	-0.287	0.349	-0.234	0.465	-0.152	0.281	-0.379	0.355	0.884*	0.341	0.364	0.345	-0.053	0.235	0.031	0.298	-0.010	0.212	-0.177	0.305
Feeding	-0.102	0.313	-0.125	0.443	-0.138	0.279	-0.555	0.356	-0.201	0.369	-0.438	0.381	0.013	0.232	-0.109	0.294	0.119	0.214	0.497	0.327
Transfer	-0.063	0.292	-0.697	0.388	-0.447	0.260	-0.604	0.324	-0.268	0.347	-0.604	0.355	0.002	0.214	-0.587*	0.252	-0.283	0.188	-0.377	0.265
Mobility	-0.708*	0.346	-1.252*	0.393	0.256	0.198	0.151	0.250	-0.339	0.297	-0.137	0.305	-0.214	0.185	-0.294	0.217	-0.172	0.156	-0.072	0.215
Dressing	0.002	0.238	0.088	0.398	-0.829*	0.209	-2.286*	0.326	-0.478	0.284	-0.976*	0.304	0.031	0.188	-0.143	0.259	-0.137	0.177	-0.219	0.275
Stairs	-0.593*	0.138	-1.093*	0.336	-0.126	0.113	-0.512*	0.244	-0.162	0.150	-0.611*	0.179	-0.201	0.114	-0.339	0.178	-0.032	0.111	-0.558*	0.214
Bathing	-0.165	0.265	-1.839	1.071	-0.325	0.223	-2.395*	1.119	-0.064	0.284	-1.108*	0.403	0.133	0.234	-0.291	0.408	0.233	0.230	-0.618	0.521
Bladder	0.226	0.229	0.586	0.365	0.099	0.196	0.209	0.278	0.222	0.246	0.243	0.268	0.381*	0.171	0.477*	0.239	-0.250	0.162	-0.141	0.247
Bowels	-0.183	0.379	-0.684	0.490	0.078	0.287	0.036	0.381	0.392	0.325	0.581	0.352	-0.244	0.246	0.228	0.347	-0.075	0.226	0.253	0.363
Constant	4.852	1.080	5.274	1.241	3.169*	0.779	4.687*	0.927	2.118*	0.944	3.487*	0.961	1.298*	0.587	0.971	0.746	1.552*	0.525	-0.554	0.807
Total Barthel	-0.210*	0.037	-0.442*	0.050	-0.238*	0.031	-0.512*	0.043	-0.094*	0.039	-0.292*	0.041	-0.045	0.023	-0.178*	0.029	-0.131*	0.022	-0.185*	0.031
Constant	4.949*	0.635	5.543*	0.741	4.408*	0.524	6.450*	0.622	3.021*	0.649	5.264*	0.665	1.392*	0.374	1.952*	0.434	1.603*	0.332	1.048*	0.442

* Statistically significant at 5% level

A8 – OLS Regression Results obtained when using the Admissions Dataset

Appendix Tables A5 and A6 show the results of the 9 regression models described in section 9.2.2 of chapter nine and performed using an OLS approach on the entire Admissions dataset.

Model 1

Table A2 shows a positive relationship between the overall EQ-5D tariff score and the Barthel overall score (a positive sign for the overall Barthel score coefficient). Age and gender were not significantly related to the overall EQ-5D tariff score. The coefficient of determination (R-squared) was 0.218.

Model 2

This model was the same as model (1) with the exception that age and gender were removed as independent variables. Here again as shown in Table A2, the overall Barthel score was positively related to the overall EQ-5D tariff score and the R-squared was 0.216.

Model 3

As shown in Table A2, four Barthel dimensions were statistically significant in this model. These were 'Transfer', 'Mobility', 'Dressing' and 'Stairs' and all of them were positively related to the overall EQ-5D tariff score. Age and gender were again not statistically significant in this model. The R-squared statistic was again higher than for the first two models at 0.241.

Model 4

When age and gender were excluded from model (3), the covariates that were statistically significant in the model are again significant in this model. In addition, however, 'Bladder' is also now significant. 'Transfer', 'Mobility', 'Dressing' and 'Stairs' also had the same positive signs as before while 'Bladder' was negatively related to the overall EQ-5D score. The sizes of the coefficients were similar but the coefficient of determination was slightly lower at 0.238.

Model 5

In this model, stepwise regression was used to reduce the number of covariates used in model (4). This resulted in only four variables being included in the model: 'Transfer', 'Mobility', 'Dressing' and 'Stairs'. All four covariates were statistically significant and also positively related to the overall EQ-5D tariff score as indicated in Table A2. The coefficient of determination was 0.237.

Model 6

Interaction terms were added to model (5) and the result was model (6). None of the main effects and interaction terms were statistically significant but the R-square was 0.234.

Model 7

In Table A3, ten levels of Barthel dimensions are shown to have been statistically significant in explaining the overall EQ-5D tariff score. These were the 'Major help', 'Minor help' and 'Independent' for the 'Transfer' dimension, 'Needs help' and 'Independent' for the 'Mobility' dimension and 'Needs help' and 'Independent' for

the ‘Dressing’ dimension. Others were ‘Needs help’ and ‘Independent’ levels for the ‘Stairs’ dimension and ‘Independent’ for the ‘Bathing’ dimension. The reference category for all of the categorical explanatory variables was the ‘dependent/unable’ category. All of the statistically significant variables had a positive relationship with EQ-5D. The R-square was 0.234.

Model 8

When age and gender were removed from the preceding model, all of the variables that were significant in that model are still significant. In addition, two more levels of the Barthel dimensions are now significant. These are the ‘Wheelchair independent’ and the ‘Occasional accident’ levels for the ‘Mobility’ and ‘Bladder’ dimensions, respectively. The R-squared was 0.246.

Model 9

This was a reduced version of Model (8) arrived at using stepwise regression. The results showed in Table A3 show that all the levels that were significant in model (8) are again significant in this model. However, the ‘Wheelchair independent’ level for the ‘Mobility’ dimension which was previously statistically significant is now statistically insignificant. An R-squared value of 0.242 was obtained.

Table A5: OLS model – Barthel dimensions entered as continuous variables: Admission Data

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Variables	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Total Barthel score	0.043**	0.003	0.042**	0.003								
Age	0.019	0.011			0.029	0.023						
Gender (1 = female, 0 = male)	-0.037	0.023			-0.029	0.023						
Barthel – Grooming					-0.004	0.028	-0.003	0.027				
Barthel – Toilet					0.033	0.027	0.033	0.027				
Barthel – Feeding					-0.009	0.027	-0.009	0.027				
Barthel – Transfer					0.072**	0.024	0.072**	0.024	0.075**	0.021	0.089	0.061
Barthel – Mobility					0.059**	0.020	0.059**	0.020	0.064**	0.018	0.079	0.055
Barthel – Dressing					0.084**	0.022	0.083**	0.022	0.098**	0.019	0.033	0.070
Barthel - Stairs					0.055**	0.014	0.052**	0.014	0.060**	0.013	0.059	0.067
Barthel – Bathing					0.056	0.029	0.058*	0.029				
Barthel – Bladder					-0.036	0.020	-0.038	0.020				
Barthel – Bowels					0.016	0.028	0.015	0.028				
Barthel - Transfer x Toilet											-0.014	0.029

Barthel - Mobility x Toilet											0.016	0.025
Barthel - Dressing x Toilet											0.021	0.031
Barthel - Stairs x Toilet											0.004	0.031
Constant	-0.244**	0.065	-0.205**	0.040	-0.104	0.074	-0.066	0.064	-0.110**	0.037	-0.099*	0.044
R-Squared		0.218		0.216		0.241		0.238		0.237		0.234

Dependent Variable; EQ-5D * Statistically significant at 5% level; ** Statistically significant at 1% level

Table A6: OLS model – Barthel dimensions entered as categorical variables: Admission Data

	Model 7		Model 8		Model 9	
Variables	Coeff.	SE	Coeff.	SE	Coeff.	SE
Age	0.024	0.013				
Gender (1 = female, 0 = male)	-0.032	0.026				
Grooming - independent	0.019	0.033	-0.008	0.028		
Toilet – needs help	0.007	0.064	0.037	0.055		
Toilet – independent	0.088	0.068	0.082	0.059		
Toilet – dependent (Reference category)						
Feeding – needs help	0.189	0.119	0.167	0.101		
Feeding - independent	0.164	0.117	0.143	0.100		
Feeding – unable (Reference category)						
Transfer – major help	0.340**	0.093	0.346**	0.082	0.308**	0.079
Transfer – minor help	0.246*	0.096	0.305**	0.086	0.288**	0.080
Transfer – independent	0.275**	0.098	0.366**	0.089	0.367**	0.082
Transfer – unable (Reference category)						
Mobility – wheelchair independent	0.107	0.096	0.067*	0.080	0.060	0.078
Mobility – needs help	0.172**	0.065	0.128**	0.057	0.129*	0.052
Mobility – independent	0.237**	0.070	0.189*	0.062	0.197**	0.057
Mobility – immobile (Reference category)						
Dressing – needs help	0.092*	0.045	0.097*	0.039	0.100**	0.037
Dressing – independent	0.150**	0.054	0.175**	0.046	0.202**	0.041
Dressing – dependent (Reference category)						
Stairs – needs help	0.071*	0.031	0.090**	0.028	0.092**	0.027
Stairs – independent	0.092**	0.033	0.101**	0.029	0.120**	0.026
Stairs – unable (Reference category)						

Bathing – independent	0.072*	0.034	0.065*	0.030		
Bladder – occasional accident	0.105	0.062	-0.099*	0.052		
Bladder – continent	0.092	0.059	-0.109	0.049		
Bladder – incontinent (Reference category)						
Bowels – occasional accident	0.029	0.093	-0.021	0.081		
Bowels – continent	0.054	0.087	0.000	0.075		
Bowels – incontinent (Reference category)						
Constant	0.125	0.143	-0.069	0.117	-0.267**	0.067
R-Squared		0.234		0.246		0.242

Dependent Variable; EQ-5D * Statistically significant at 5% level

** Statistically significant at 1% level

A9 – OLS Regression Results obtained when using the split Admissions Dataset

Appendix Tables A7 and A8 show the results of the 9 regression models described in section 9.2.2 of chapter nine and performed using an OLS approach on the split Admissions dataset.

Model 1

Table A4 shows a positive relationship between the overall EQ-5D tariff score and the Barthel overall score (a positive sign for the overall Barthel score coefficient). Age was significantly related to the overall EQ-5D tariff score while Gender was not. The coefficient of determination (R-squared) was 0.217.

Model 2

This model was the same as model (1) with the exception that age and gender were removed as independent variables. Here again as shown in Table A4, the overall Barthel score was positively related to the overall EQ-5D tariff score and the R-squared was 0.212.

Model 3

As shown in Table A4, four Barthel dimensions were statistically significant in this model. These were 'Mobility', 'Dressing', 'Stairs' and 'Bathing' and all of them were positively related to the overall EQ-5D tariff score. Age and gender were again not statistically significant in this model. The R-squared statistic was again higher than for the first two models at 0.224.

Model 4

When age and gender were excluded from model (3), the covariates that were statistically significant in the model are again significant in this model. All statistically significant variables also had the same positive signs as before. The sizes of the coefficients were similar but the coefficient of determination was slightly lower at 0.220.

Model 5

In this model, stepwise regression was used to reduce the number of covariates used in model (4). This resulted in only four variables being included in the model: 'Transfer', 'Mobility', 'Dressing' and 'Stairs'. All four covariates were statistically significant and also positively related to the overall EQ-5D tariff score as indicated in Table A4. The coefficient of determination was 0.218.

Model 6

Interaction terms were added to model (5) and the result was model (6). None of the main effects and interaction terms were statistically significant but the R-square was 0.212.

Model 7

In Table A5, ten levels of Barthel dimensions are shown to have been statistically significant in explaining the overall EQ-5D tariff score. These were the 'Major help', 'Minor help' and 'Independent' for the 'Transfer' dimension, 'Needs help' and 'Independent' for the 'Mobility' dimension and 'Needs help' and 'Independent' for the 'Dressing' dimension. Others were 'Needs help' and 'Independent' levels for the

‘Stairs’ dimension and ‘Independent’ for the ‘Bathing’ dimension. The reference category for all of the categorical explanatory variables was the ‘dependent/unable’ category. All of the statistically significant variables had a positive relationship with EQ-5D. The R-square was 0.234.

Model 8

When age and gender were removed from the preceding model, all of the variables that were significant in that model are still significant. The R-squared was 0.230.

Model 9

This was a reduced version of Model (8) arrived at using stepwise regression. The results showed in Table A5 show that all the levels that were significant in model (8) are again significant in this model. An R-squared value of 0.223 was obtained.

Table A7: OLS model – Barthel dimensions entered as continuous variables: Split Admission Data

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Variables	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Total Barthel score	0.043**	0.003	0.042**	0.003								
Age	0.027*	0.013			0.023	0.013						
Gender (1 = female, 0 = male)	-0.030	0.026			-0.024	0.026						
Barthel – Grooming					0.023	0.032	0.022	0.032				
Barthel – Toilet					0.045	0.030	0.043	0.030				
Barthel – Feeding					-0.008	0.031	-0.009	0.031				
Barthel – Transfer					0.042	0.026	0.041	0.026	0.054*	0.024	0.058	0.070
Barthel – Mobility					0.067**	0.022	0.068**	0.022	0.076**	0.021	0.080	0.064
Barthel – Dressing					0.069**	0.025	0.069**	0.025	0.099**	0.021	0.037	0.082
Barthel – Stairs					0.045**	0.016	0.042**	0.016	0.055**	0.015	0.052	0.076
Barthel – Bathing					0.069*	0.033	0.070*	0.033				
Barthel – Bladder					-0.021	0.024	-0.025	0.024				
Barthel – Bowels					0.032	0.032	0.033	0.032				

Barthel - Transfer x Toilet											-0.011	0.033
Barthel - Mobility x Toilet											0.021	0.030
Barthel - Dressing x Toilet											0.018	0.035
Barthel - Stairs x Toilet											0.004	0.035
Constant	-0.271**	0.065	-0.195**	0.046	-0.141	0.084	-0.075	0.071	-0.081	0.042	-0.056	0.050
R-Squared		0.217		0.212		0.224		0.220		0.218		0.212

Dependent Variable; EQ-5D * Statistically significant at 5% level; ** Statistically significant at 1% level

Table A8: OLS model – Barthel dimensions entered as categorical variables: Split Admission Data

	Model 7		Model 8		Model 9	
Variables	Coeff.	SE	Coeff.	SE	Coeff.	SE
Age	0.024	0.013				
Gender (1 = female, 0 = male)	-0.032	0.026				
Grooming - independent	0.019	0.033	0.019	0.033		
Toilet – needs help	0.007	0.064	0.016	0.064		
Toilet – independent	0.088	0.068	0.088	0.068		
Toilet – dependent (Reference category)						
Feeding – needs help	0.189	0.119	0.194	0.118		
Feeding - independent	0.164	0.117	0.172	0.117		
Feeding – unable (Reference category)						
Transfer – major help	0.340**	0.093	0.338**	0.093	0.301**	0.090
Transfer – minor help	0.246*	0.096	0.241*	0.096	0.236*	0.091
Transfer – independent	0.275**	0.098	0.272**	0.098	0.304**	0.093
Transfer – unable (Reference category)						
Mobility – wheelchair independent	0.107	0.096	0.119	0.094	0.088	0.090
Mobility – needs help	0.172**	0.065	0.171**	0.065	0.148*	0.060
Mobility – independent	0.237**	0.070	0.241**	0.070	0.236**	0.064
Mobility – immobile (Reference category)						
Dressing – needs help	0.092*	0.045	0.094*	0.045	0.104*	0.042
Dressing – independent	0.150**	0.054	0.149**	0.053	0.200**	0.046
Dressing – dependent (Reference category)						
Stairs – needs help	0.071*	0.031	0.065*	0.031	0.070*	0.030
Stairs – independent	0.092**	0.033	0.086**	0.033	0.113**	0.030

Stairs – unable (Reference category)						
Bathing – independent	0.072*	0.034	0.074*	0.034		
Bladder – occasional accident	0.105	0.062	-0.110	0.062		
Bladder – continent	0.092	0.059	-0.101	0.059		
Bladder – incontinent (Reference category)						
Bowels – occasional accident	0.029	0.093	0.036	0.093		
Bowels – continent	0.054	0.087	0.059	0.087		
Bowels – incontinent (Reference category)						
Constant	0.125	0.143	-0.065	0.134	-0.237*	0.076
R-Squared		0.234		0.230		0.223

Dependent Variable; EQ-5D

* Statistically significant at 5% level

** Statistically significant at 1% level

A10-Plots of Residuals (Admission & Split Admissions datasets)

Figure A1: Plot of Residuals - Model 1 (OLS), Admission dataset

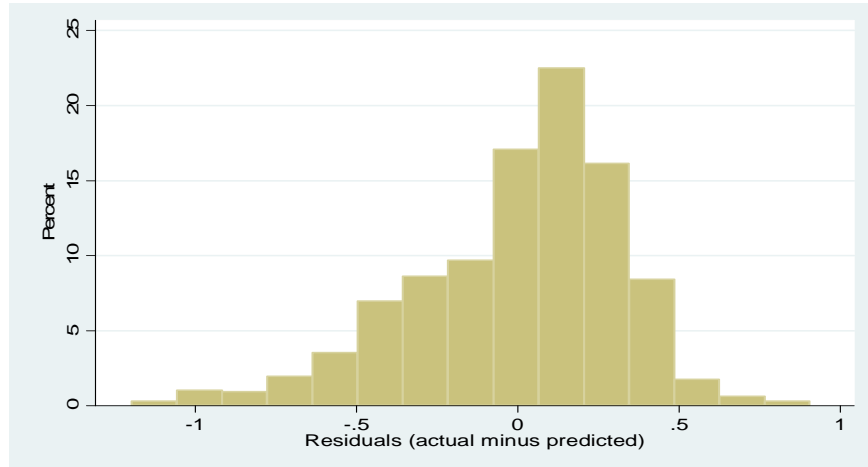


Figure A2: Plot of Residuals - Model 1 (CLAD), Admission dataset

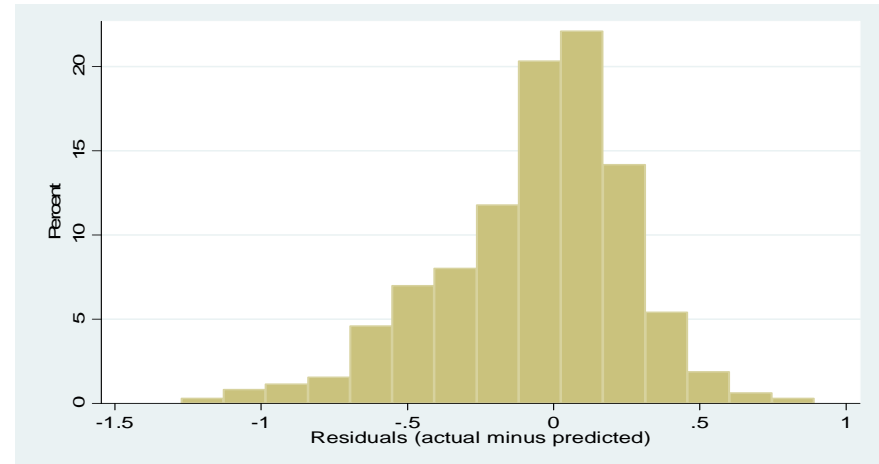


Figure A3: Plot of Residuals - Model 2 (OLS), Admission dataset

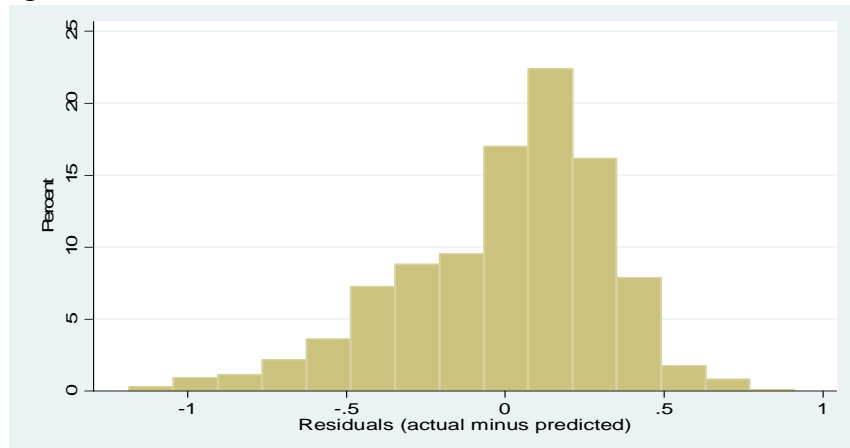


Figure A4: Plot of Residuals - Model 2 (CLAD), Admission dataset

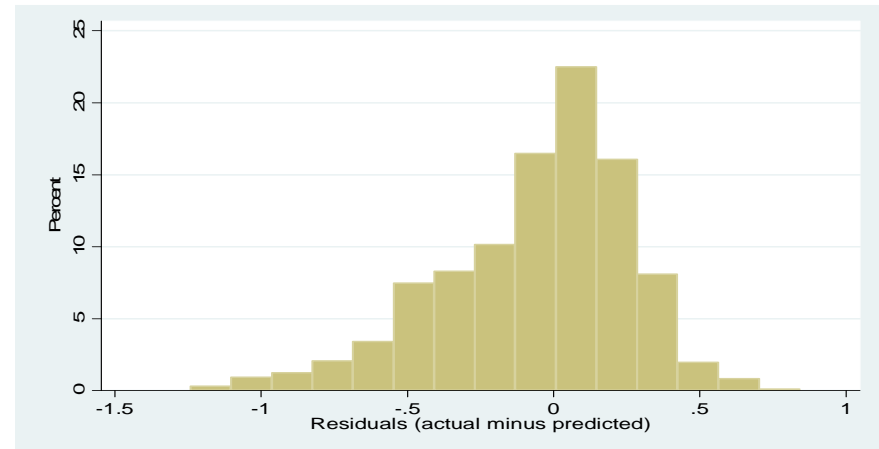


Figure A5: Plot of Residuals - Model 3 (OLS), Admission dataset

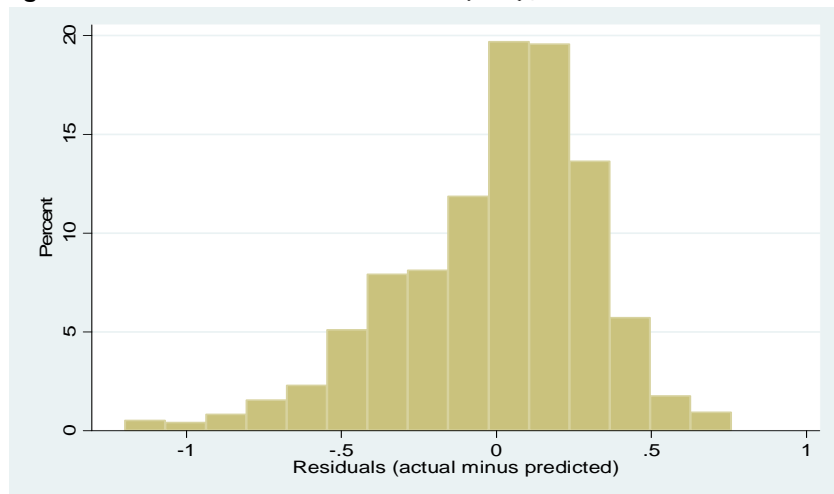


Figure A6: Plot of Residuals - Model 3 (CLAD), Admission dataset

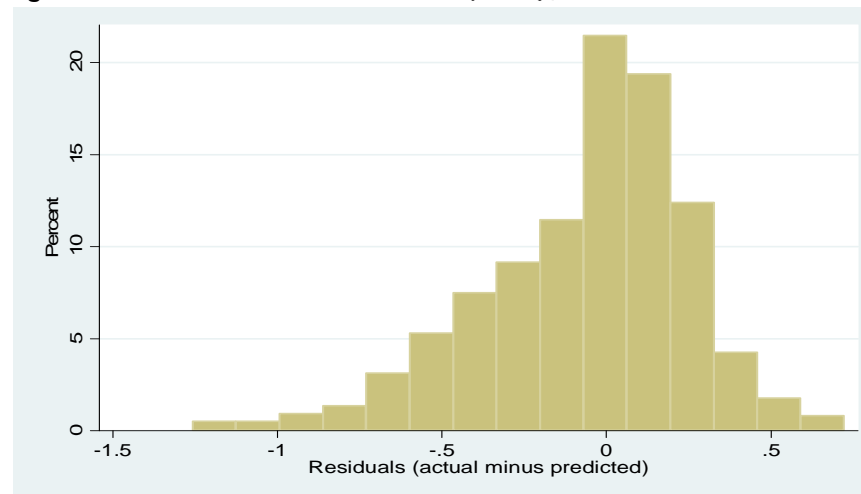


Figure A7: Plot of Residuals - Model 4 (OLS), Admission dataset

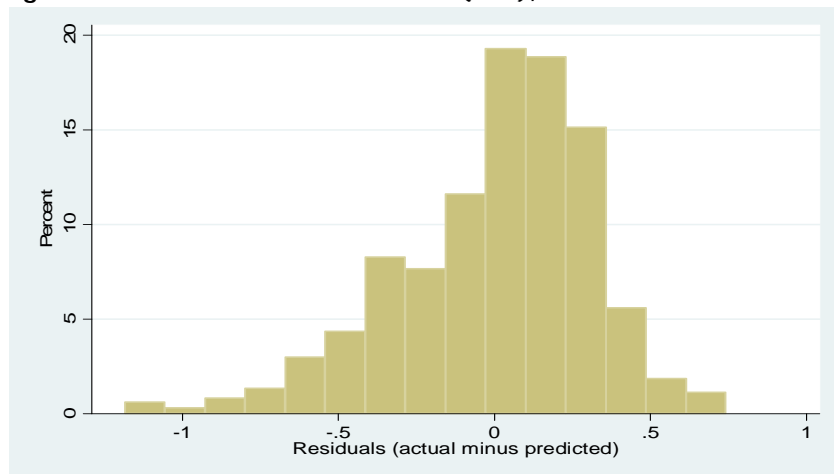


Figure A8: Plot of Residuals - Model 4 (CLAD), Admission dataset

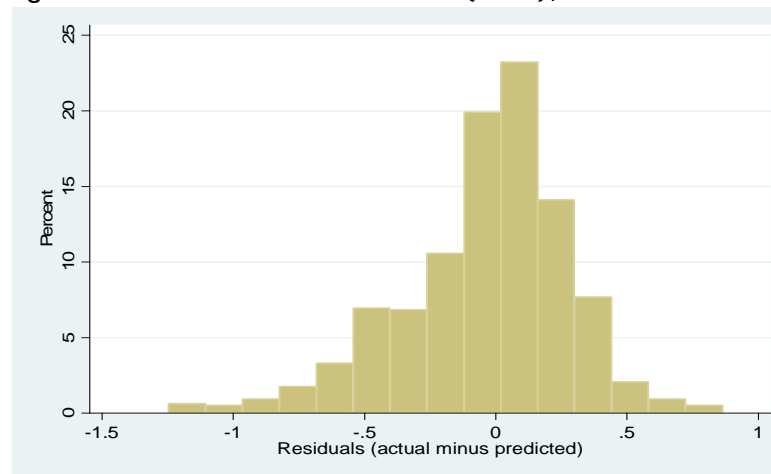


Figure A9: Plot of Residuals - Model 5 (OLS), Admission dataset

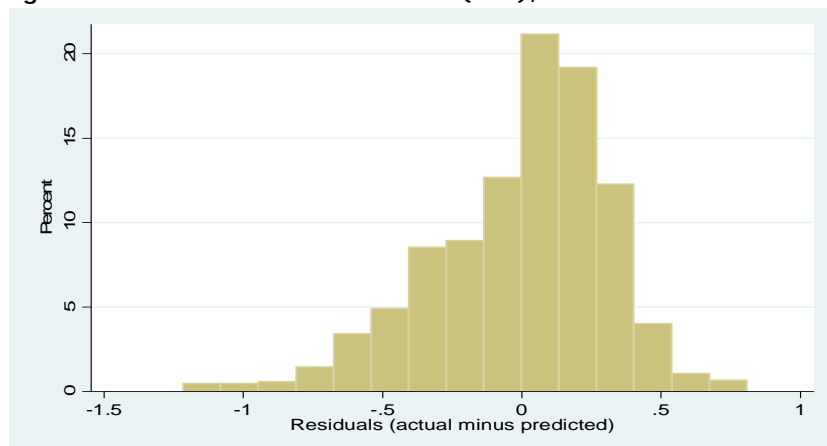


Figure A10: Plot of Residuals - Model 5 (CLAD), Admission dataset

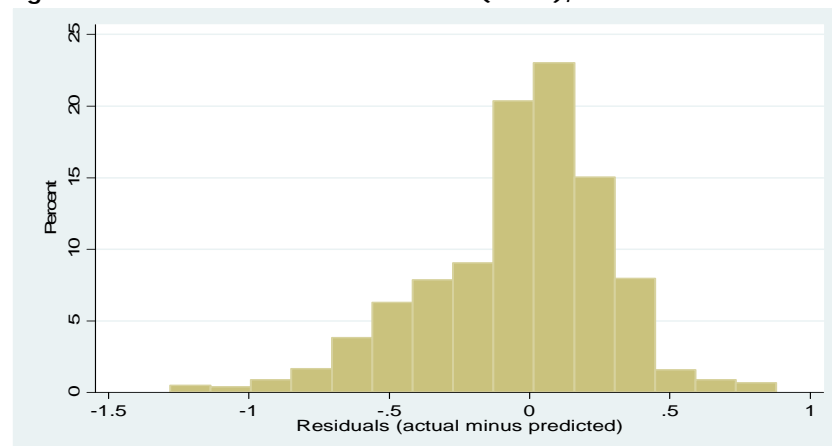


Figure A11: Plot of Residuals - Model 6 (OLS), Admission dataset

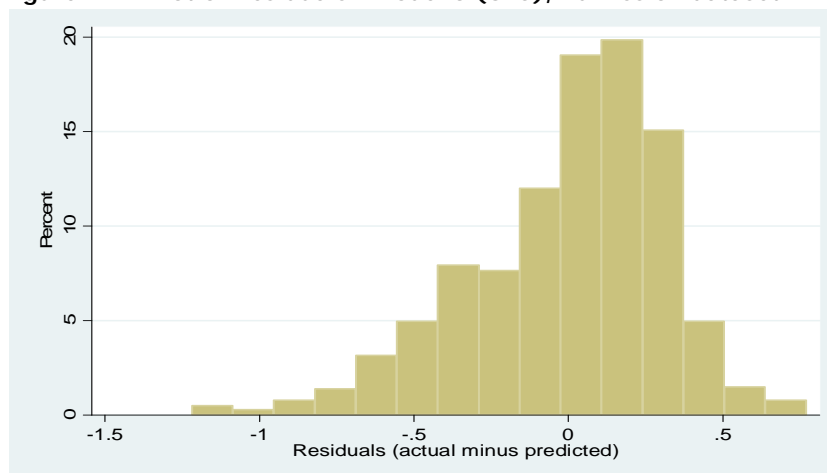


Figure A12: Plot of Residuals - Model 6 (CLAD), Admission dataset

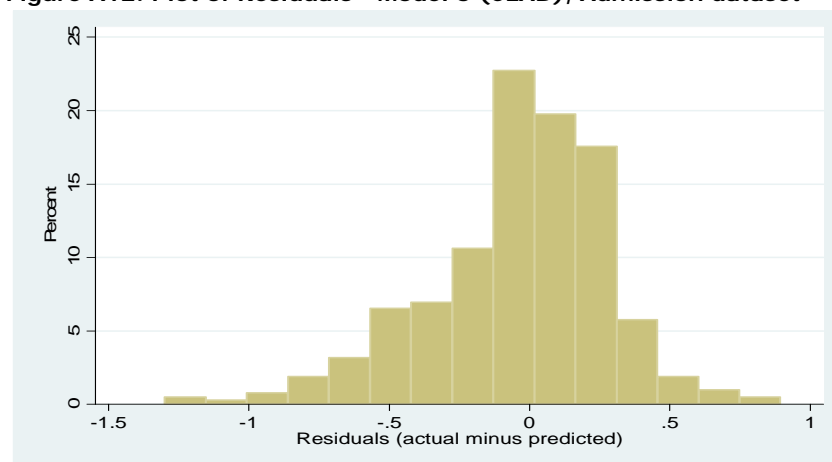


Figure A13: Plot of Residuals - Model 7 (OLS), Admission dataset

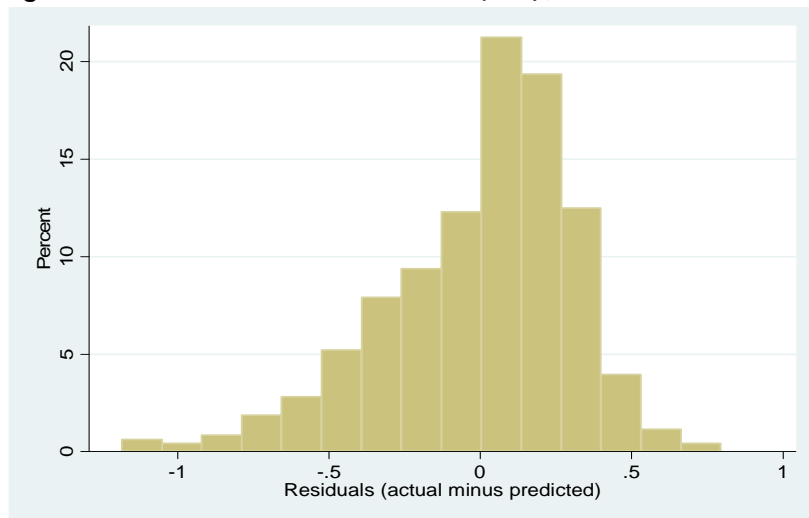


Figure A14: Plot of Residuals - Model 7 (CLAD), Admission dataset

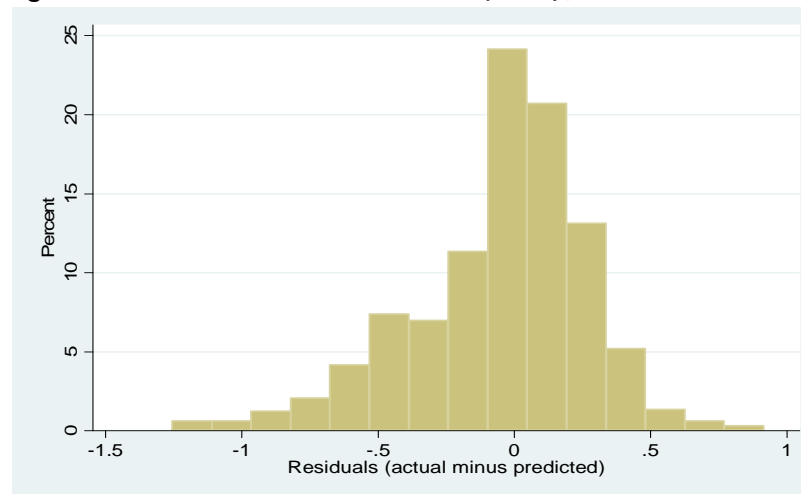


Figure A15: Plot of Residuals - Model 8 (OLS), Admission dataset

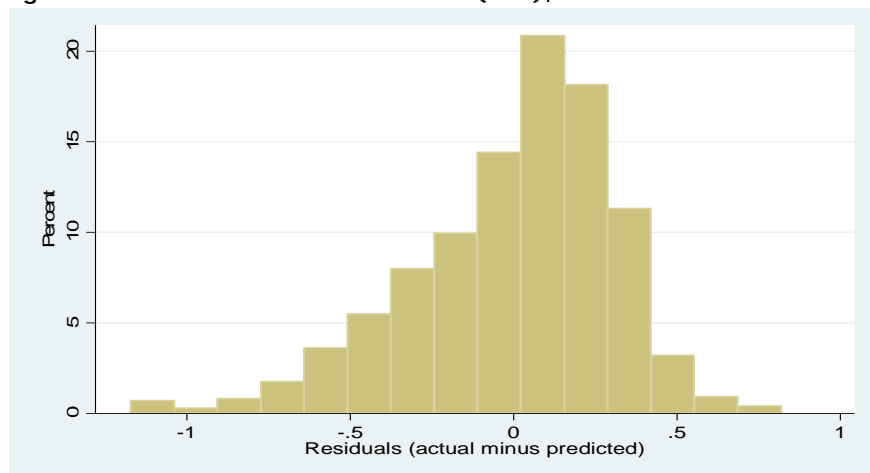


Figure A16: Plot of Residuals - Model 8 (CLAD), Admission dataset

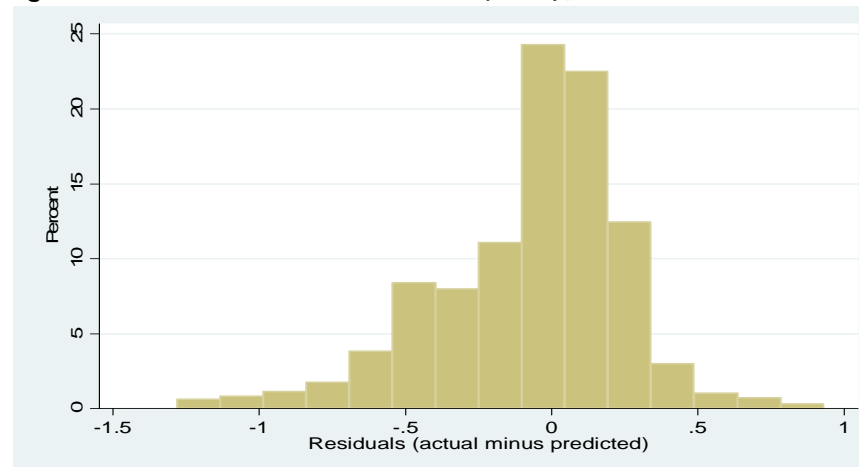


Figure A17: Plot of Residuals - Model 9 (OLS), Admission dataset

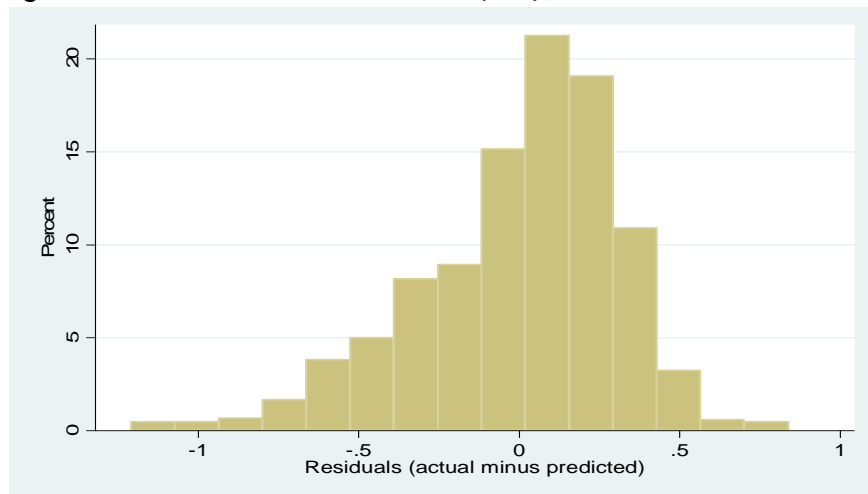


Figure A18: Plot of Residuals - Model 9 (CLAD), Admission dataset

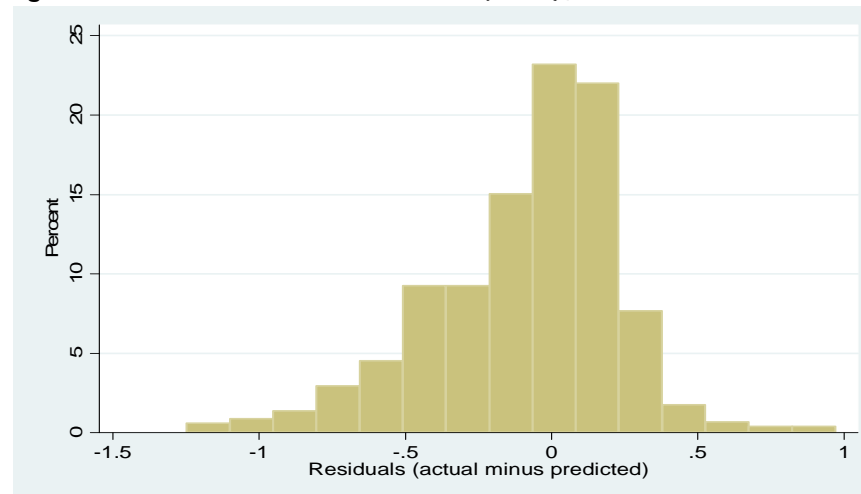


Figure A19: Plot of Residuals - Model 10* (Multinomial_Montercarlo), Admission dataset

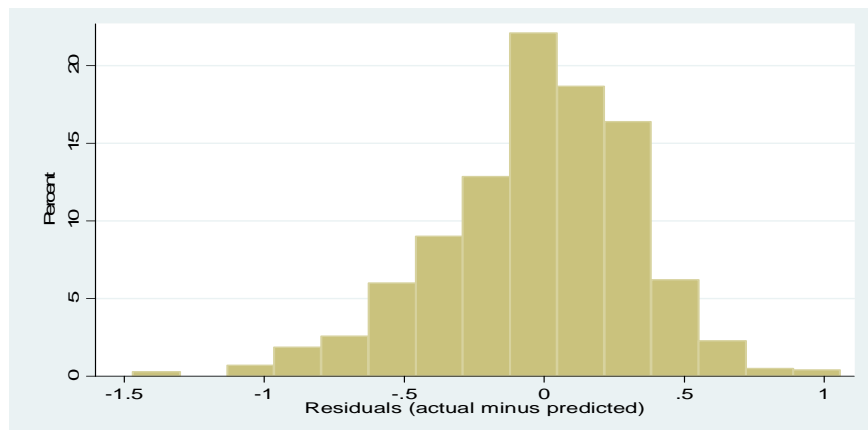


Figure A20: Plot of Residuals - Model 10# (Multinomial_Crude Method), Admission dataset

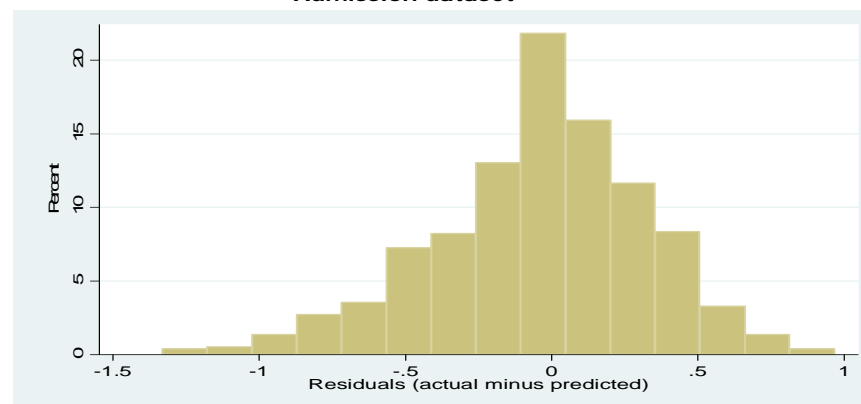


Figure A21: Plot of Residuals - Model 11* (Multinomial_Montercarlo), Admission dataset

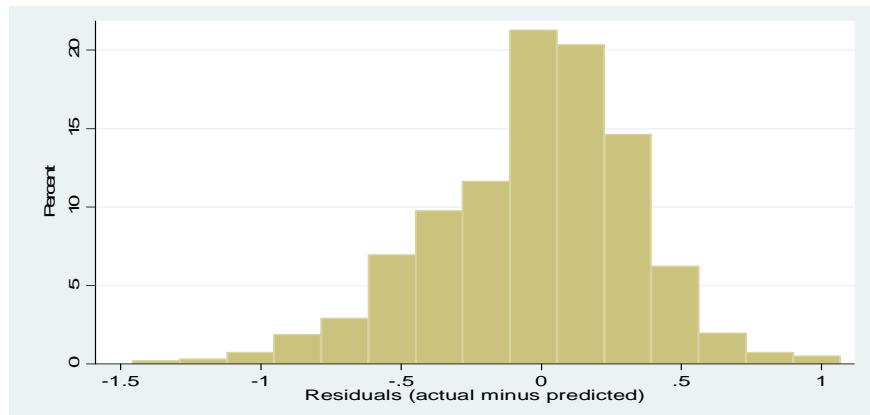


Figure A22: Plot of Residuals - Model 11* (Multinomial_Montercarlo), Admission dataset

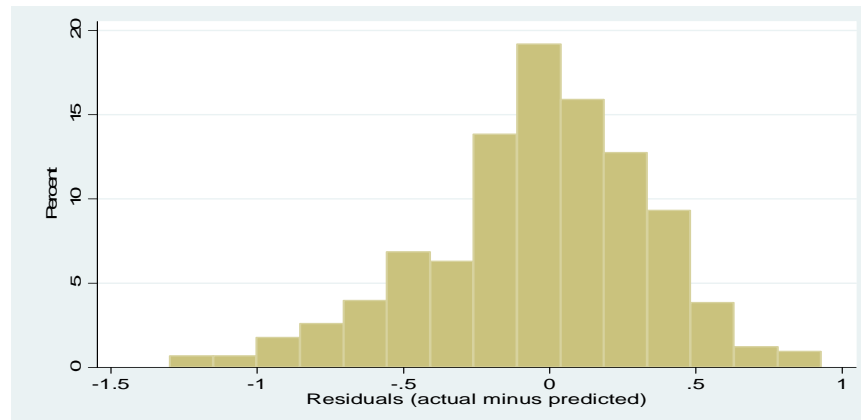


Figure A23: Plot of Residuals - Model 1 (OLS), Split Admission dataset [within sample]

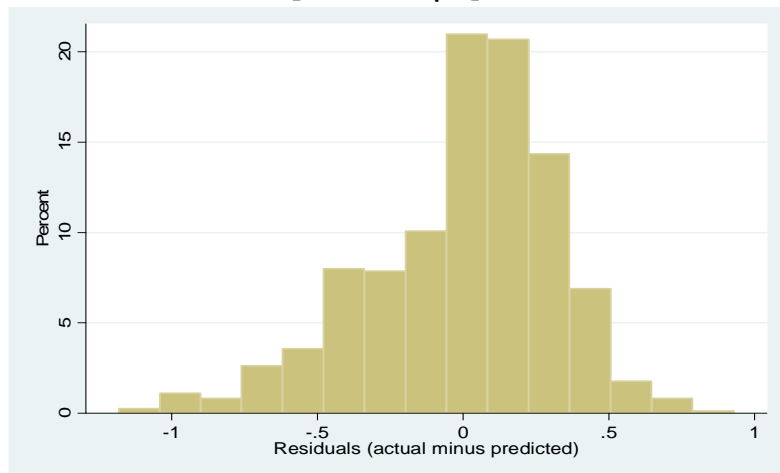


Figure A24: Plot of Residuals - Model 1 (CLAD), Split Admission dataset [within sample]

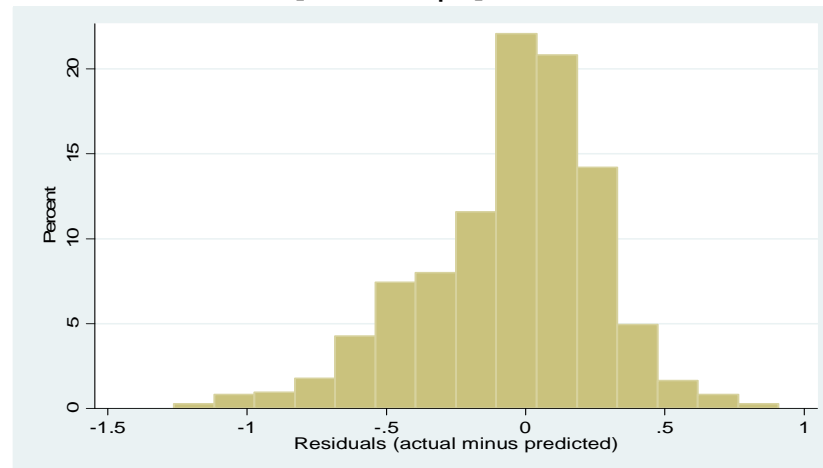


Figure A25: Plot of Residuals - Model 2 (OLS), Split Admission dataset [within sample]

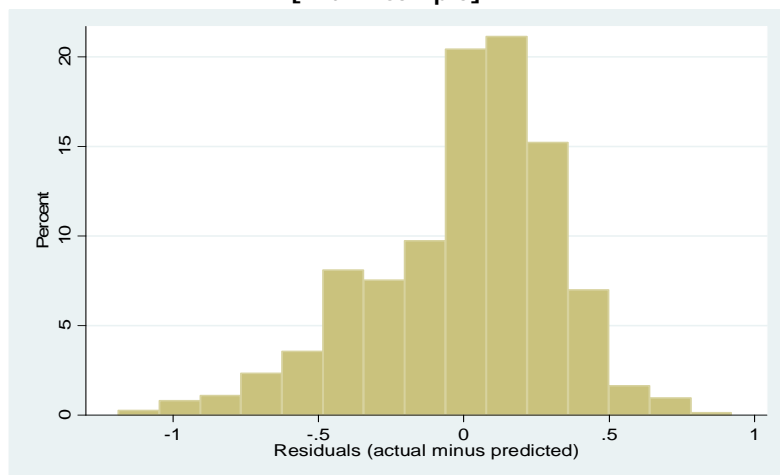


Figure A26: Plot of Residuals - Model 2 (CLAD), Split Admission dataset [within sample]

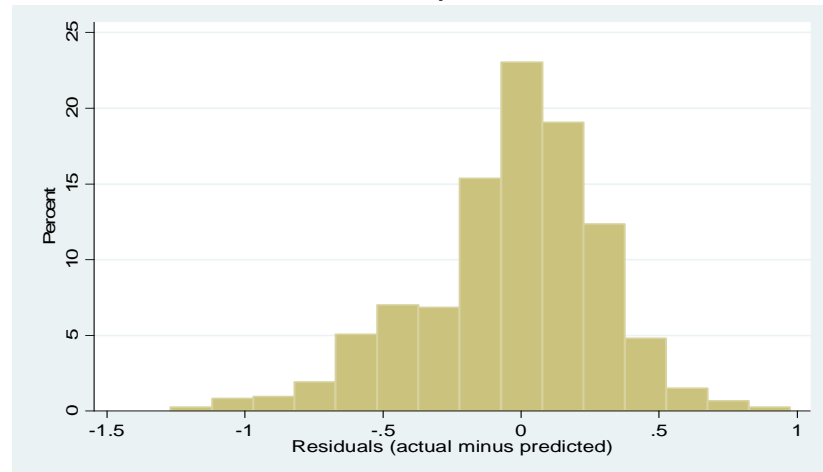


Figure A27: Plot of Residuals - Model 3 (OLS), Split Admission dataset [within sample]

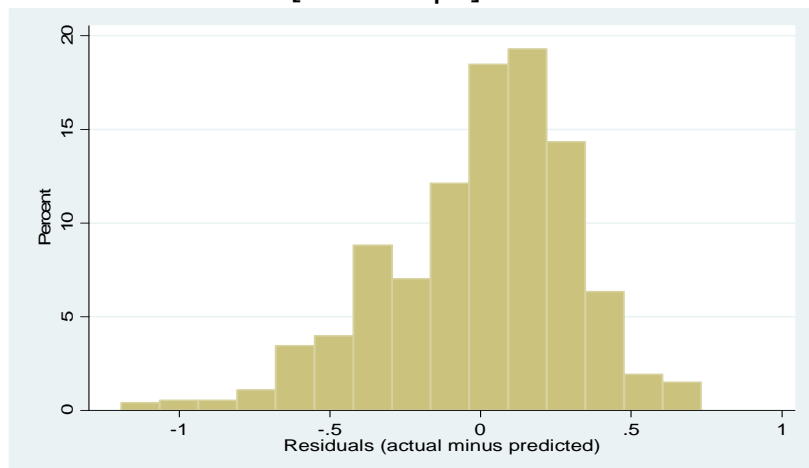


Figure A28: Plot of Residuals - Model 3 (CLAD), Split Admission dataset [within sample]

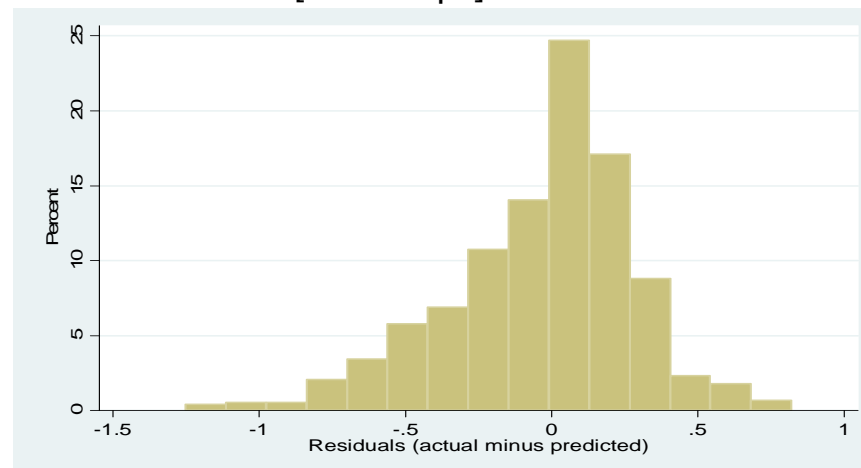


Figure A29: Plot of Residuals - Model 4 (OLS) - Split Admission dataset [within sample]

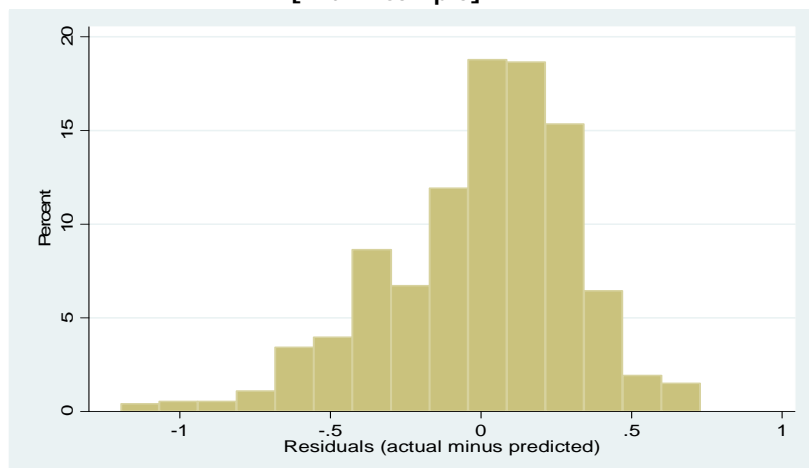


Figure A30: Plot of Residuals - Model 4 (CLAD) - Split Admission dataset [within sample]

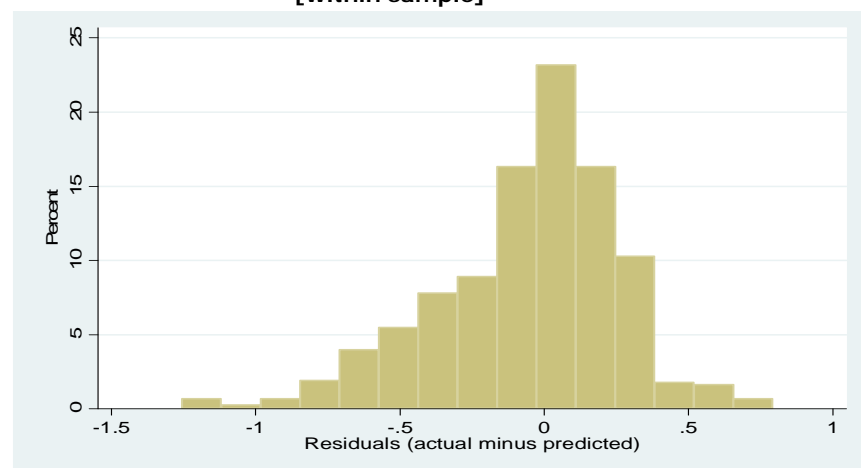


Figure A40: Plot of Residuals - Model 5 (OLS), Split Admission dataset [within sample]

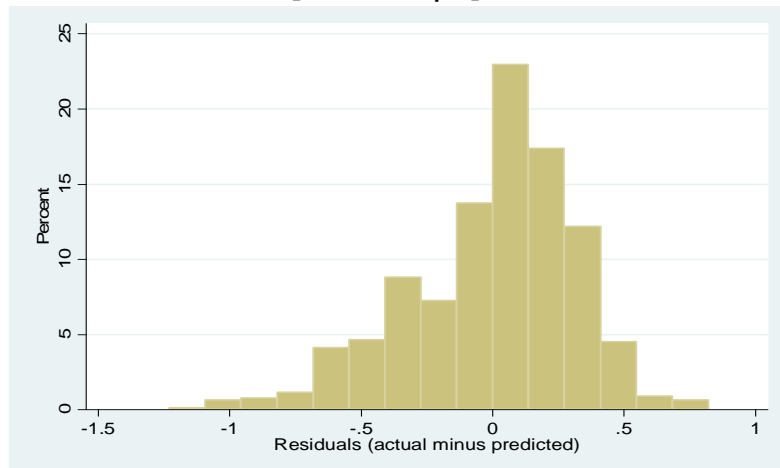


Figure A41: Plot of Residuals - Model 5 (CLAD), Split Admission dataset [within sample]

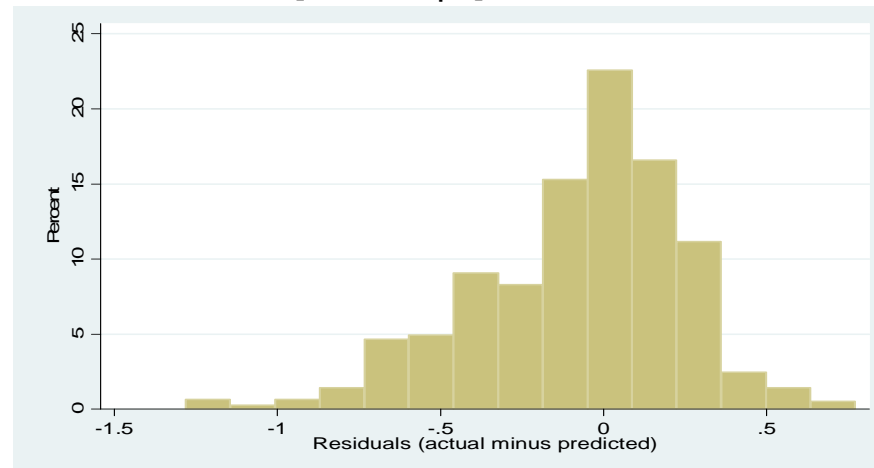


Figure A42: Plot of Residuals - Model 6 (OLS), Split Admission dataset [within sample]

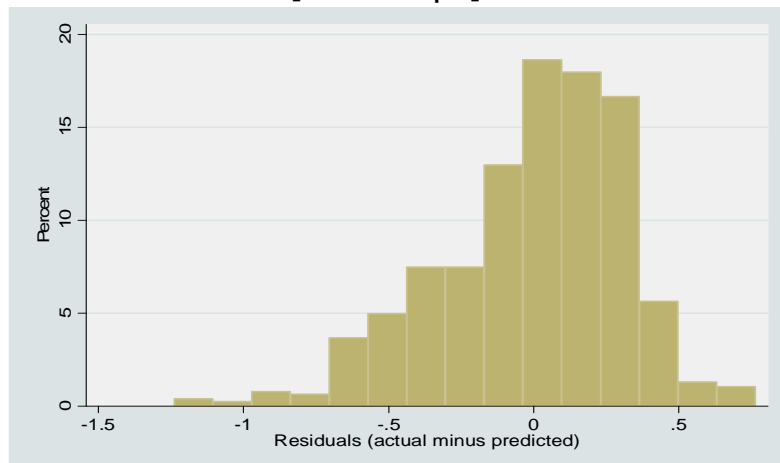


Figure A43: Plot of Residuals - Model 6 (CLAD), Split Admission dataset [within sample]

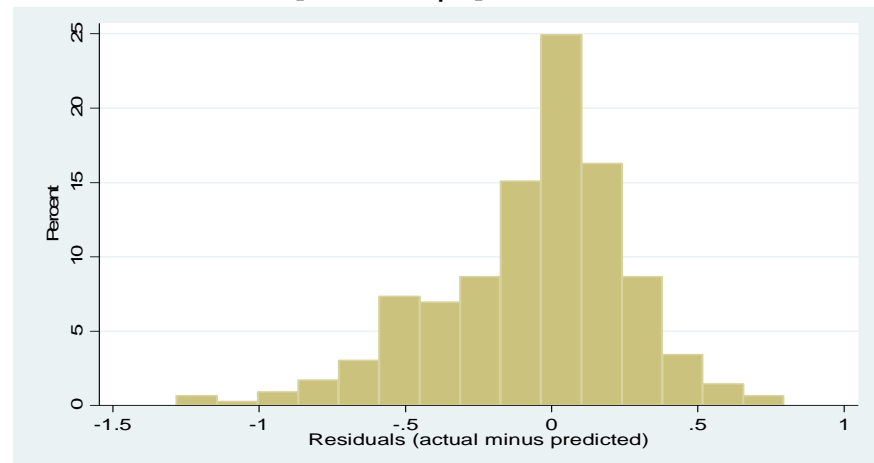


Figure A44: Plot of Residuals - Model 7 (OLS), Split Admission dataset [within sample]

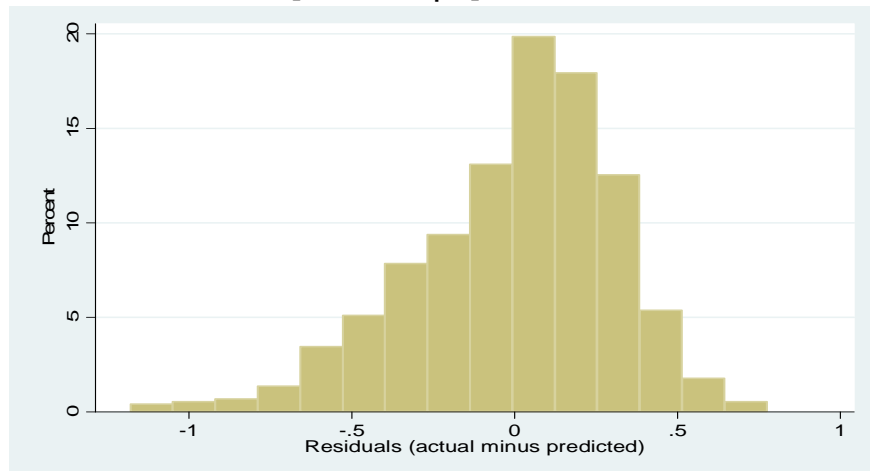


Figure A45: Plot of Residuals - Model 7 (CLAD), Split Admission dataset [within sample]

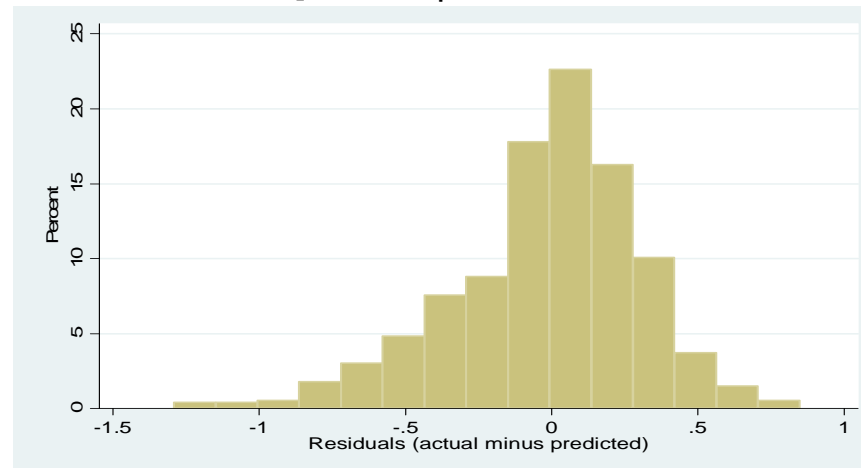


Figure A46: Plot of Residuals - Model 8 (OLS), Split Admission dataset [within sample]

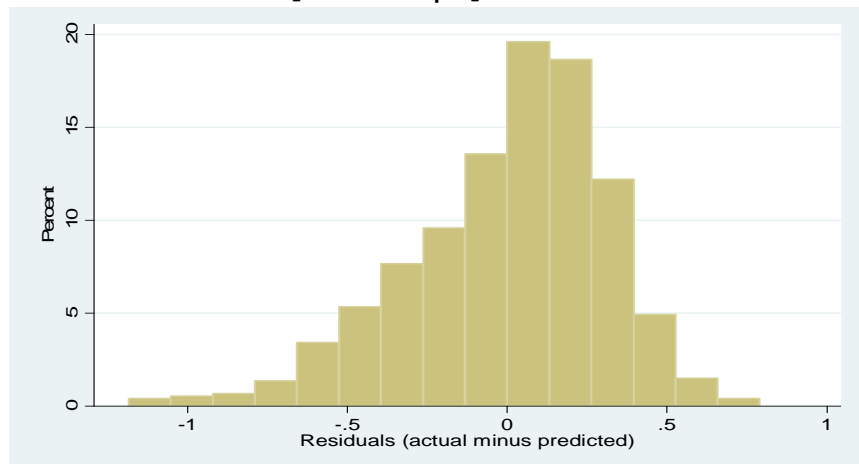


Figure A47: Plot of Residuals - Model 8 (CLAD), Split Admission dataset [within sample]

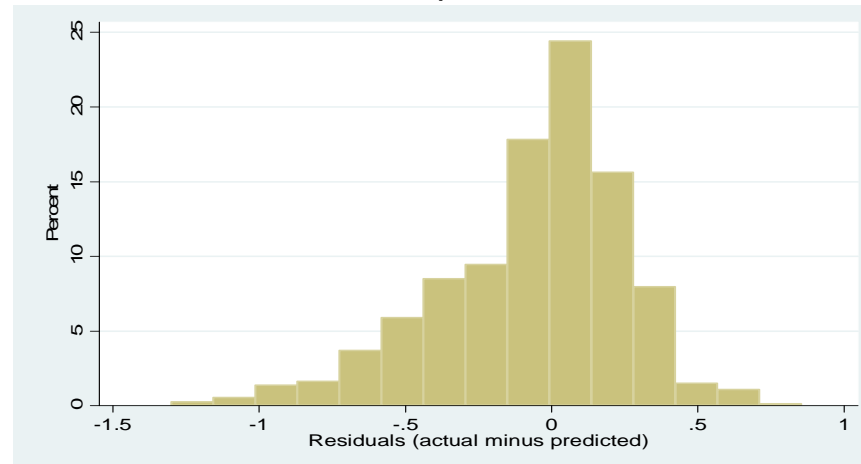


Figure A48: Plot of Residuals - Model 9 (OLS), Split Admission dataset [within sample]

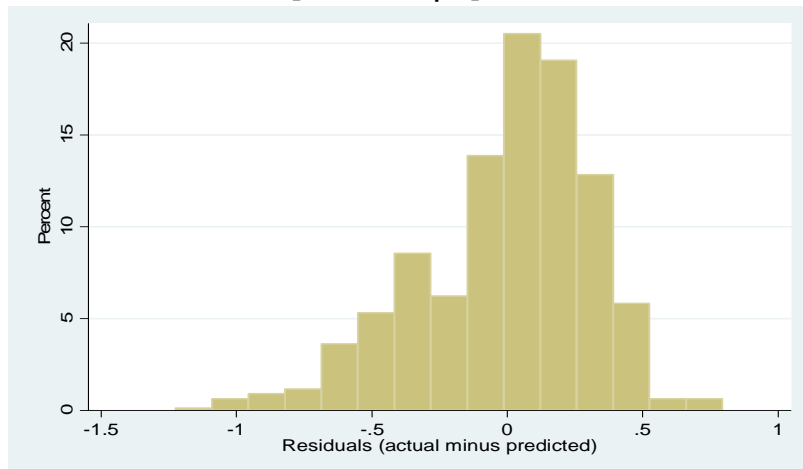


Figure A49: Plot of Residuals - Model 9 (CLAD), Split Admission dataset [within sample]

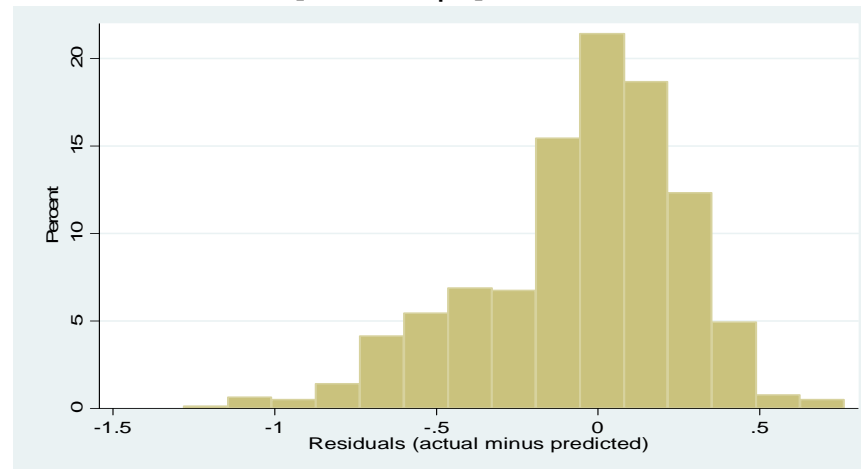


Figure A50: Plot of Residuals - Model 10* (Multinomial_Montercarlo), Split Admission dataset [within sample]

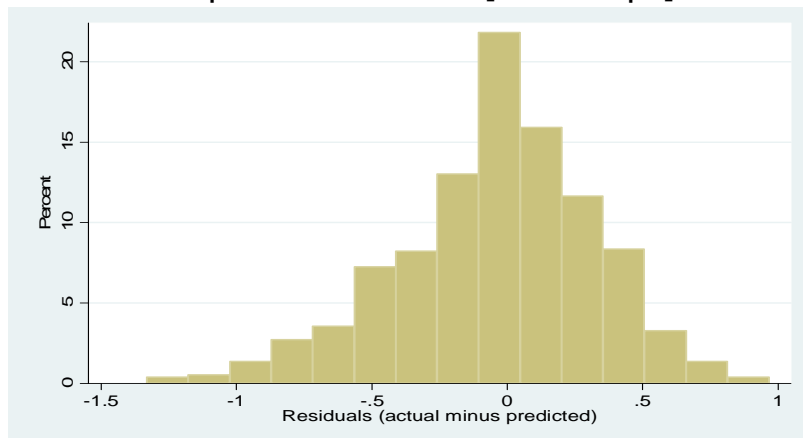
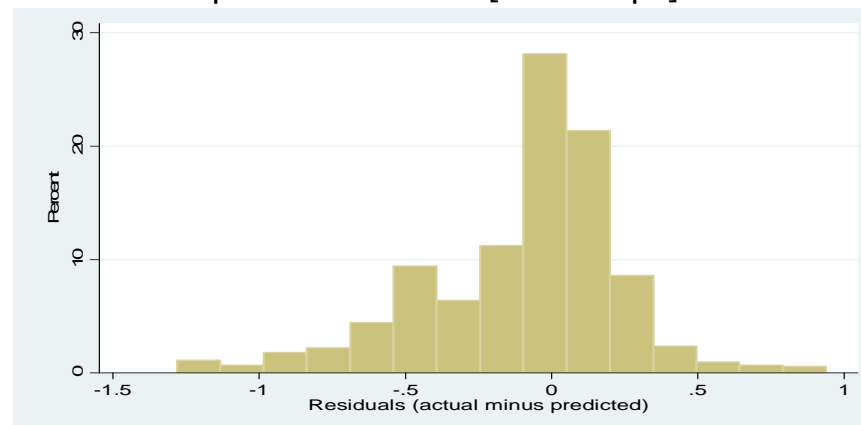
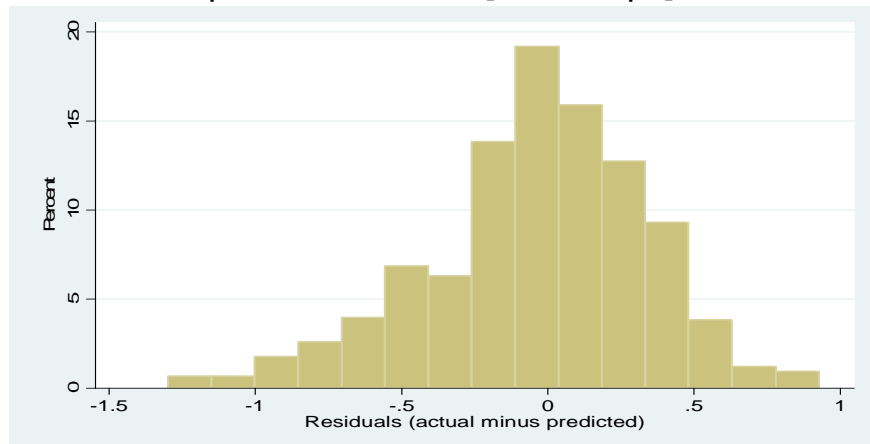


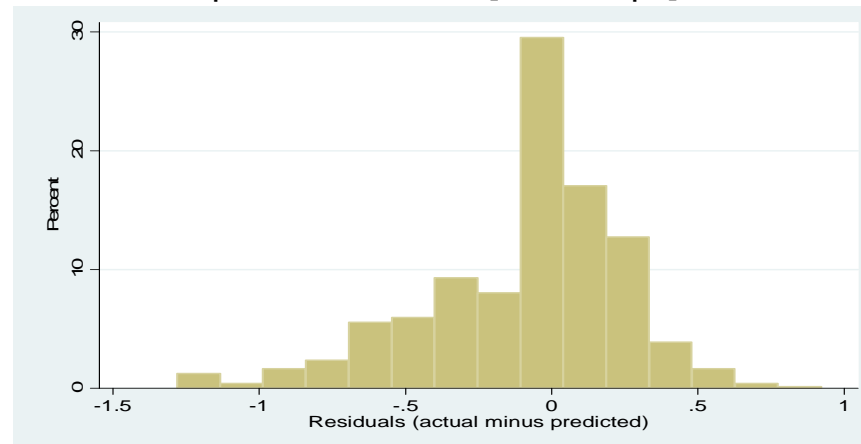
Figure A51: Plot of Residuals - Model 10# (Multinomial_Crude Method), Split Admission dataset [within sample]



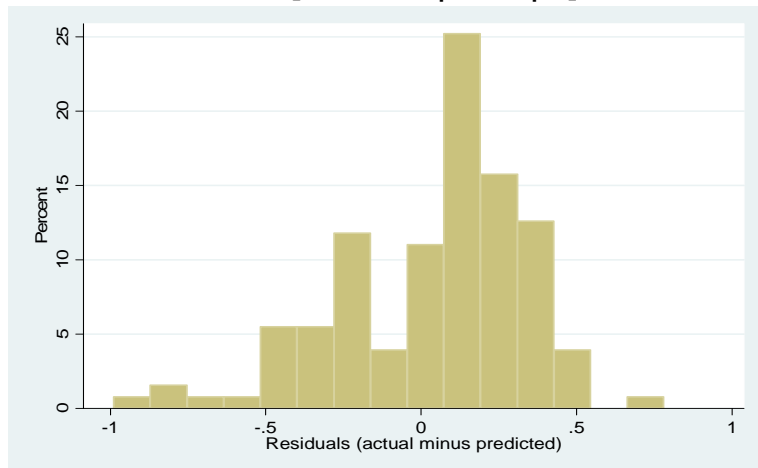
**Figure A52: Plot of Residuals - Model 11* (Multinomial_Montercarlo),
Split Admission dataset [within sample]**



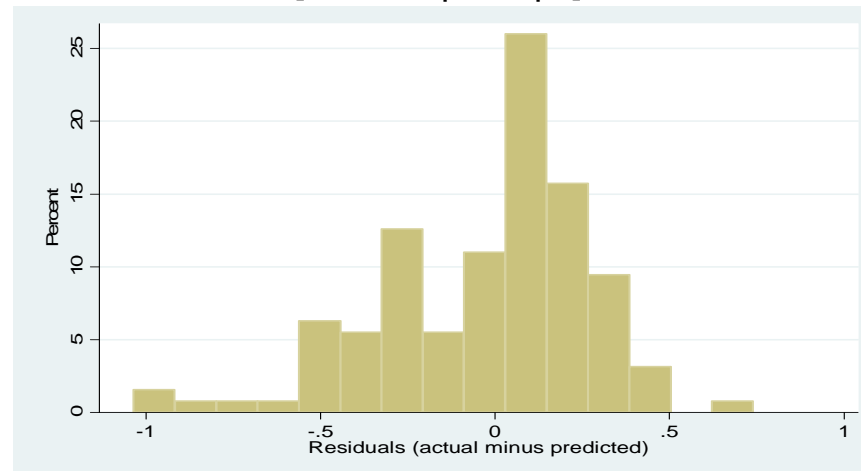
**Figure A53: Plot of Residuals - Model 11# (Multinomial_Crude Method),
Split Admission dataset [within sample]**



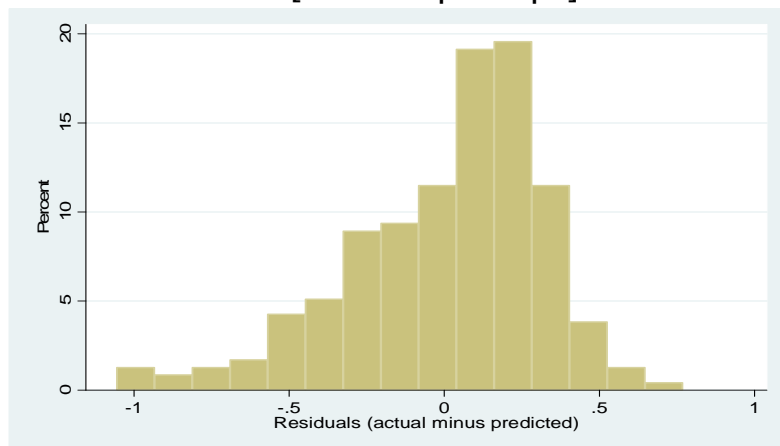
**Figure A54: Plot of Residuals - Model 1 (OLS), Split Admission dataset
[Out-of-sample sample]**



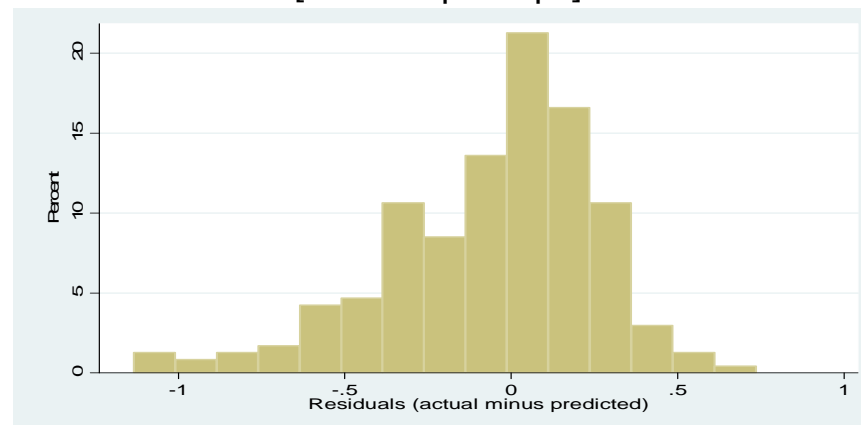
**Figure A55: Plot of Residuals - Model 1 (CLAD), Split Admission dataset
[Out-of-sample sample]**



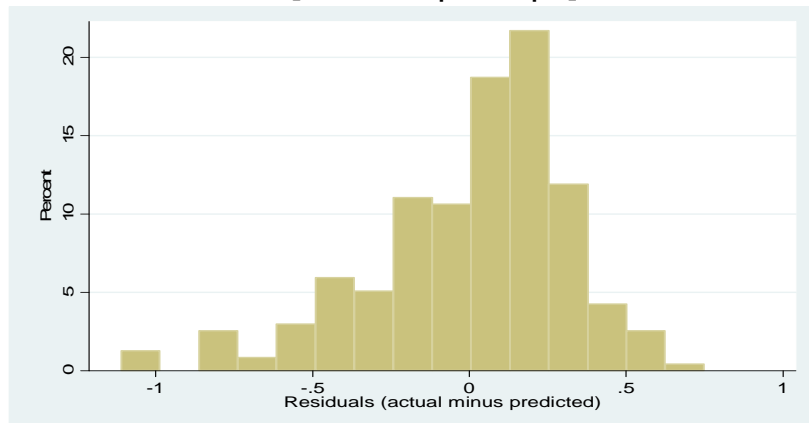
**Figure A56: Plot of Residuals - Model 2 (OLS), Split Admission dataset
[Out-of-sample sample]**



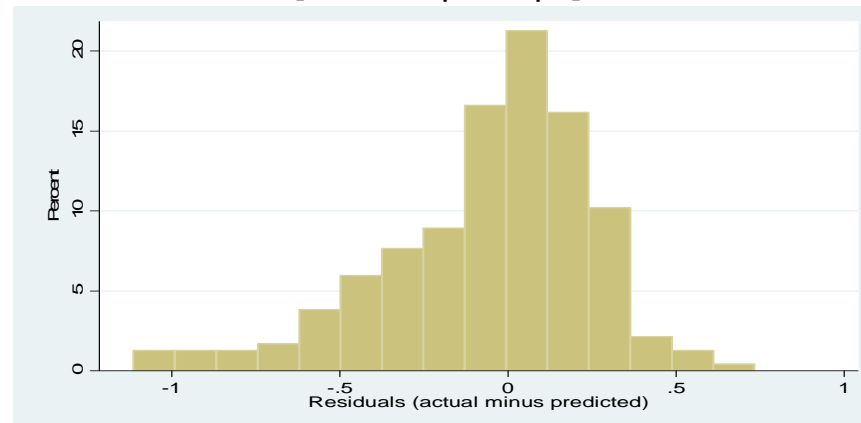
**Figure A57: Plot of Residuals - Model 2 (CLAD), Split Admission dataset
[Out-of-sample sample]**



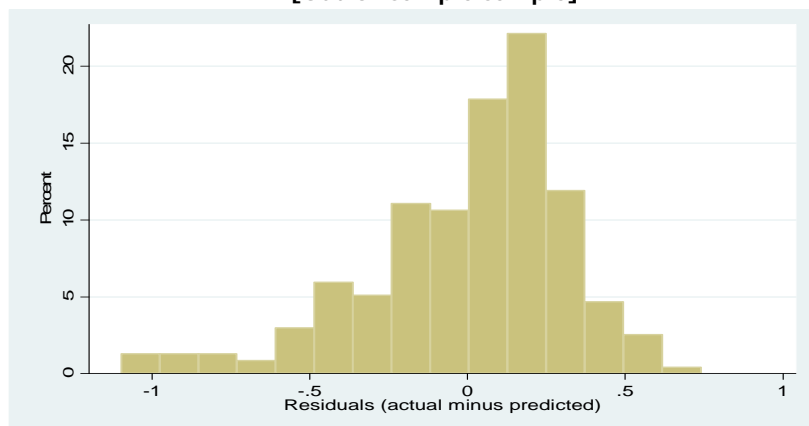
**Figure A58: Plot of Residuals - Model 3 (OLS), Split Admission dataset
[Out-of-sample sample]**



**Figure A59: Plot of Residuals - Model 3 (CLAD), Split Admission dataset
[Out-of-sample sample]**



**Figure A60: Plot of Residuals - Model 4 (OLS), Split Admission dataset
[Out-of-sample sample]**



**Figure A61: Plot of Residuals - Model 4 (CLAD), Split Admission dataset
[Out-of-sample sample]**

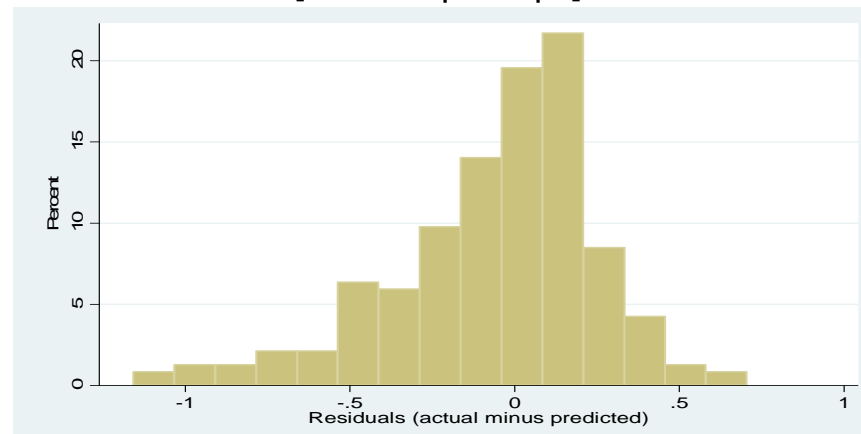


Figure A62: Plot of Residuals - Model 5 (OLS), Split Admission dataset [Out-of-sample sample]

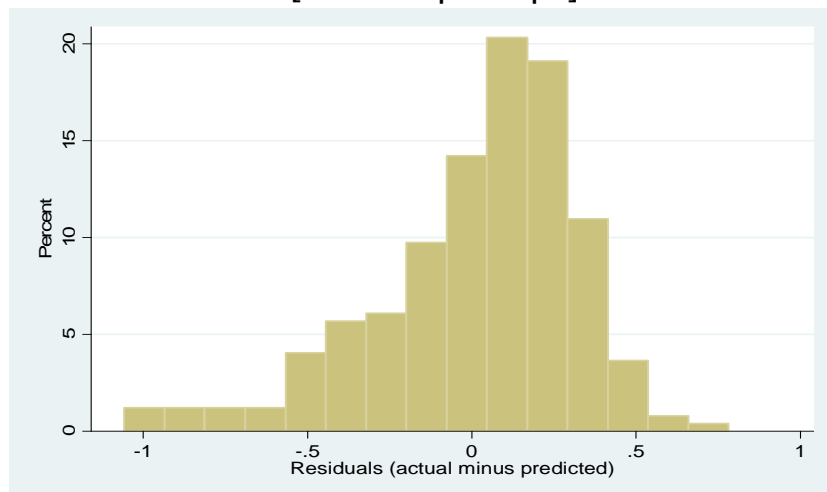


Figure A63: Plot of Residuals - Model 5 (CLAD), Split Admission dataset [Out-of-sample sample]

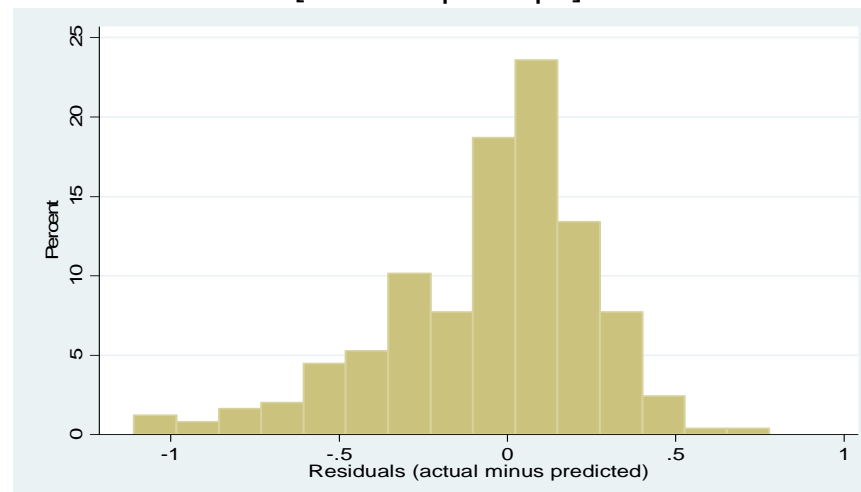


Figure A64: Plot of Residuals - Model 6 (OLS), Split Admission dataset [Out-of-sample sample]

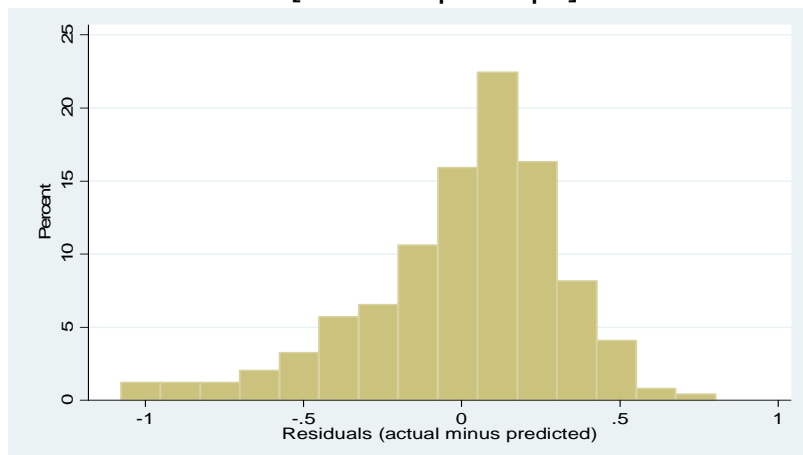
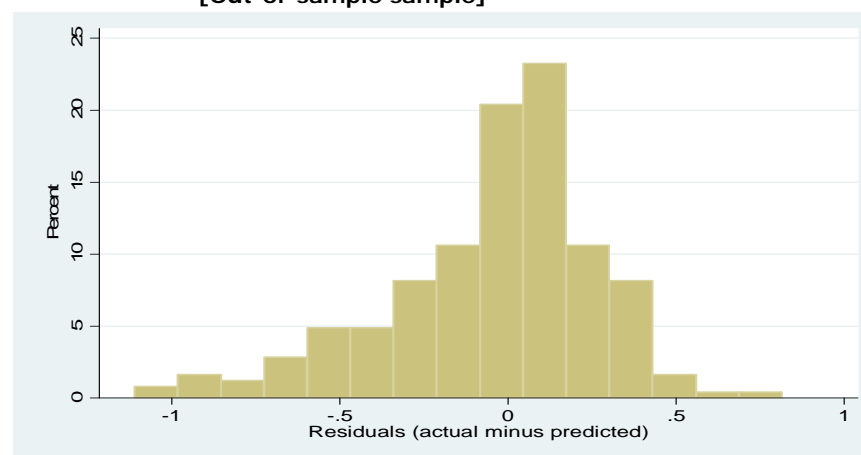
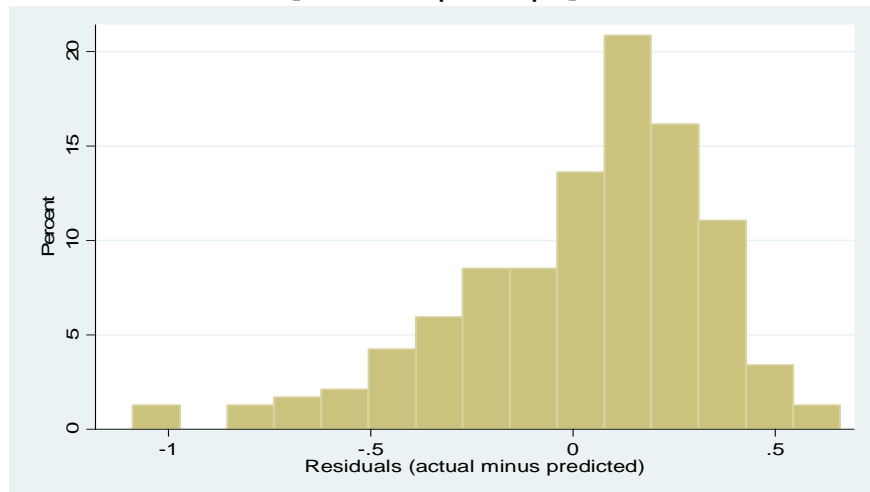


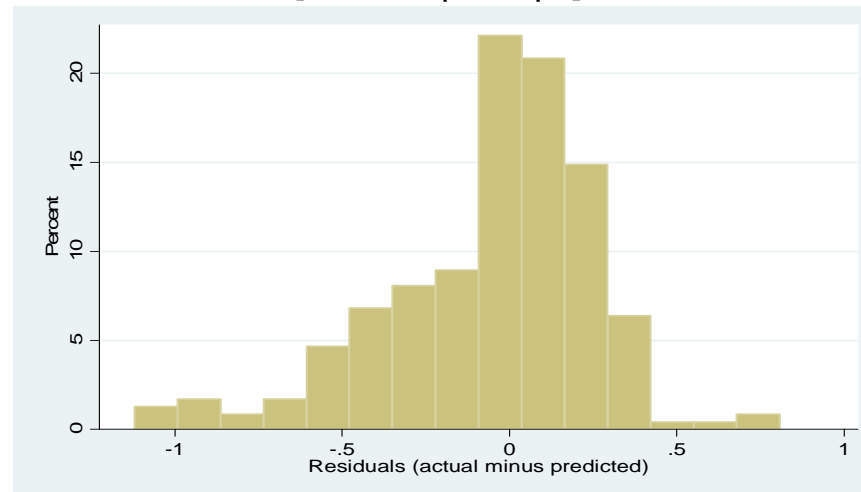
Figure A65: Plot of Residuals - Model 6 (CLAD), Split Admission dataset [Out-of-sample sample]



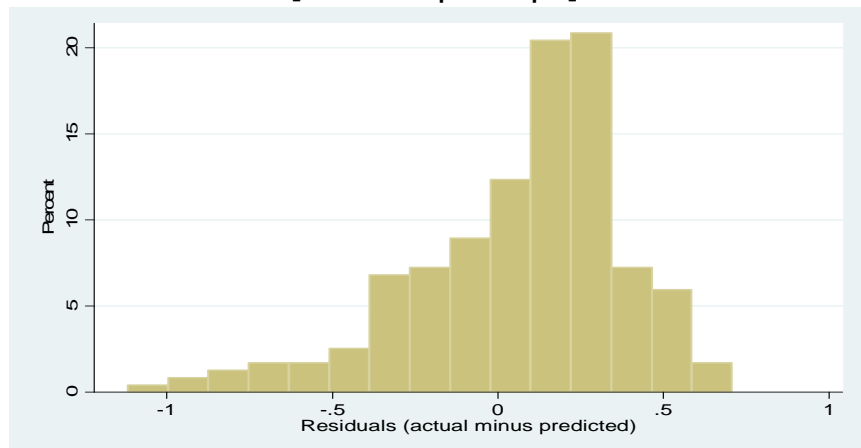
**Figure A66: Plot of Residuals - Model 7 (OLS), Split Admission dataset
[Out-of-sample sample]**



**Figure A67: Plot of Residuals - Model 7 (CLAD), Split Admission dataset
[Out-of-sample sample]**



**Figure A68: Plot of Residuals - Model 8 (OLS), Split Admission dataset
[Out-of-sample sample]**



**Figure A69: Plot of Residuals - Model 8 (CLAD), Split Admission dataset
[Out-of-sample sample]**

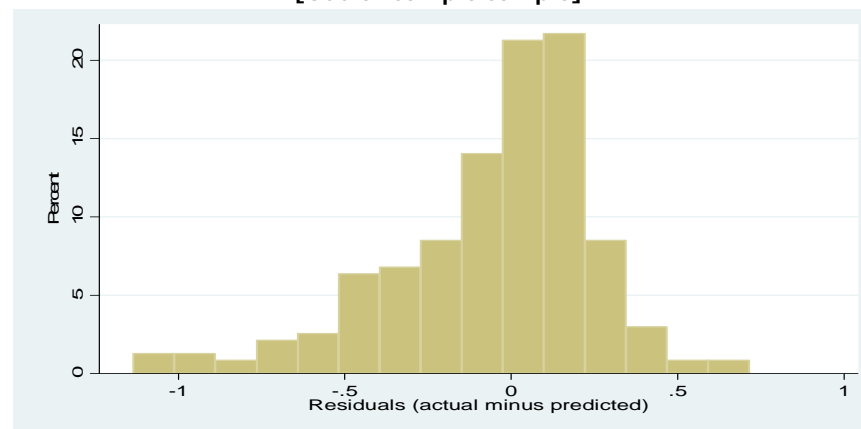


Figure A70: Plot of Residuals - Model 9 (OLS), Split Admission dataset [Out-of-sample sample]

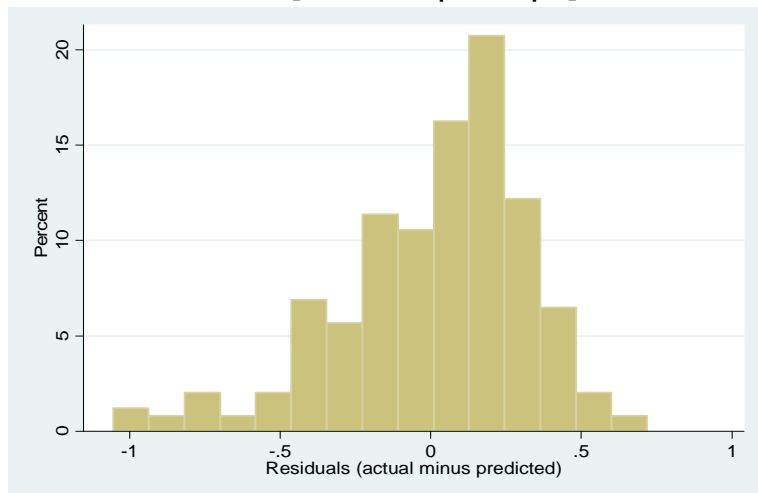


Figure A71: Plot of Residuals - Model 9 (CLAD), Split Admission dataset [Out-of-sample sample]

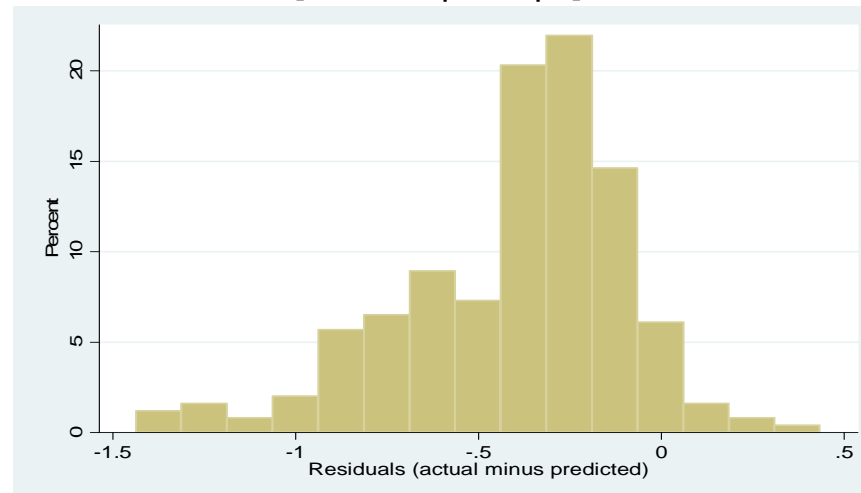


Figure A72: Plot of Residuals - Model 10* (Multinomial_Montercarlo), Split Admission dataset [out-of-sample]

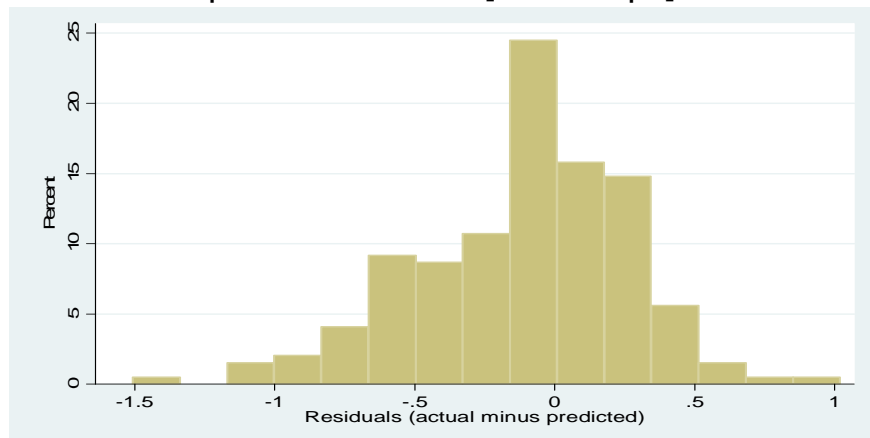
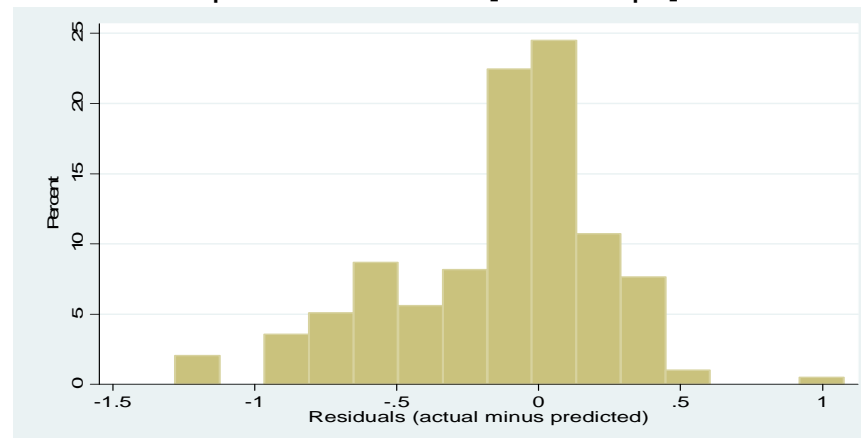
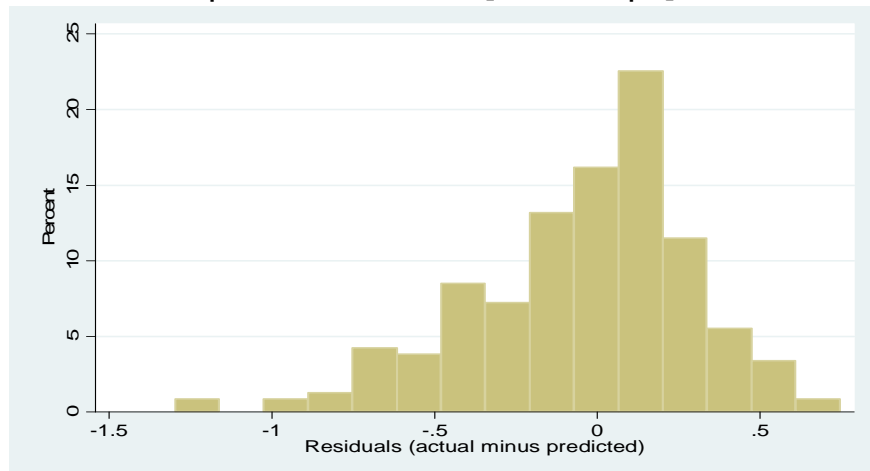


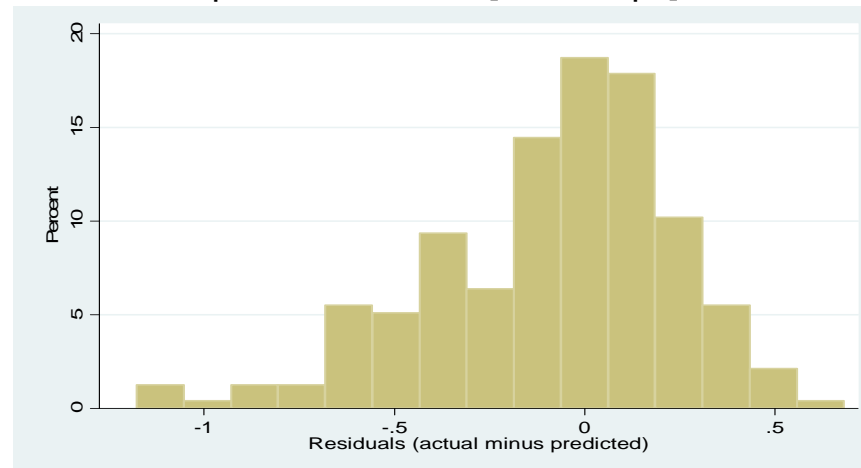
Figure A73: Plot of Residuals - Model 10# (Multinomial_Crude Method), Split Admission dataset [out-of-sample]



**Figure A74: Plot of Residuals - Model 11* (Multinomial_Montercarlo),
Split Admission dataset [out-of-sample]**



**Figure A75: Plot of Residuals - Model 11* (Multinomial_Crude Method),
Split Admission dataset [out-of-sample]**



REFERENCES

- Acock, A. 1997, "Working with missing data", *Family Science Review*, vol. 1, pp. 76-102.
- Altman, D. 1991, *Practical statistics for medical research*, Second edn, Chapman & Hall, London.
- Amemiya, T. 2008, "Tobit models: a survey", *Journal of Econometrics*, vol. 24, pp. 3-61.
- Anon 2000, *Collins English Dictionary: 21st Century Edition*, 5th edn, HarperCollins Publishers Limited, Aylesbury.
- Arabmazar, A. & Schmidt, P. 1981, "Further evidence on the robustness of the Tobit estimator to heteroskedasticity", *Journal of Econometrics*, vol. 17, no. 2, pp. 253-258.
- Audit Commission 1997, *The Coming of Age: Improving Care Services for Older People.*, Audit Commission, London.
- Audit Commission 2000, *The way to go home: rehabilitation and remedial services for older people*, Audit Commission, London.
- Austin, P. C. 2002, "A comparison of methods for analyzing health-related quality-of-life measures", *Value Health.*, vol. 5, no. 4, pp. 329-337.
- Bago-d'Uva, T. 2005, "Latent class models for use of primary care: evidence from a British panel", *Health Econ.*, vol. 14, no. 9, pp. 873-892.
- Bajekal, M., Blane, D., Grewal, I., Karlsen, S., & Nazroo, J. 2004, "Ethnic differences in influences on quality of life at older ages: a quantitative analysis", *Ageing & Society*, vol. 24, pp. 709-728.
- Baker, S. G. & Laird, N. M. 1988, "Regression analysis for categorical variables with outcome subject to nonignorable nonresponse", *Journal of the American Statistical Association*, vol. 83, pp. 62-69.
- Bansback, N., Marra, C., Tsuchiya, A., Anis, A., Guh, D., Hammond, T., & Brazier, J. 2007, "Using the health assessment questionnaire to estimate preference-based single indices in patients with rheumatoid arthritis", *Arthritis Rheum.*, vol. 57, no. 6, pp. 963-971.
- Barber, J. A. & Thompson, S. G. 2000, "Analysis of cost data in randomized trials: an application of the non-parametric bootstrap", *Stat.Med.*, vol. 19, no. 23, pp. 3219-3236.
- Barnard, J. & Meng, X. L. 1999, "Applications of multiple imputation in medical studies: from AIDS to NHANES", *Stat.Methods Med Res.*, vol. 8, no. 1, pp. 17-36.

Barofsky, I., Erickson, P., & Eberhardt, M. 2004, "Comparison of a single global item and an index of a multi-item health status measure among persons with and without diabetes in the US", *Qual.Life Res.*, vol. 13, no. 10, pp. 1671-1681.

Bartman, B. A., Rosen, M. J., Bradham, D. D., Weissman, J., Hochberg, M., & Revicki, D. A. 1998, "Relationship between health status and utility measures in older claudicants", *Qual.Life Res.*, vol. 7, no. 1, pp. 67-73.

Barton, G., Sach, T., Jenkinson, C., Avery, A., Doherty, M., & Muir, K. Mapping utility scores: How does it compare to actual trial results. Health Economics Study Group, University of East Anglia, January 2008. 2008a

Barton, G. R., Sach, T. H., Jenkinson, C., Avery, A. J., Doherty, M., & Muir, K. R. 2008b, "Do estimates of cost-utility based on the EQ-5D differ from those based on the mapping of utility scores?", *Health Qual Life Outcomes.*, vol. 6:51., p. 51.

Beale, N., Hollinghurst, S., Taylor, G., Gwynne, M., Peart, C., & Straker-Cook, D. 2005, "The costs of care in general practice: patients compared by the council tax valuation band of their home address", *Fam.Pract.*, vol. 22, no. 3, pp. 317-322.

Beaumont, J. G. & Kenealy, P. 2004, "Quality of life perceptions and social comparisons in healthy old age", *Ageing & Society*, vol. 24, pp. 755-769.

Beech, R. 2005, "Evidence on the effectiveness of intermediate care," in *Intermediate and Continuing Care: Policy and Practice*, B. Roe & R. Beech, eds., Blackwell Publishing Ltd, Oxford, pp. 106-118.

Beech, R., Russell, W., Little, R., & Sherlow-Jones, S. 2004, "An evaluation of a multidisciplinary team for intermediate care at home", *Int.J Integr.Care.*, vol. 4, p. e02.

Blackburn, C., Read, J., & Hughes, N. 2005, "Carers and the digital divide: factors affecting Internet use among carers in the UK", *Health Soc.Care Community.*, vol. 13, no. 3, pp. 201-210.

Blough, D. K., Madden, C. W., & Hornbrook, M. C. 1999, "Modeling risk using generalized linear models", *J Health Econ.*, vol. 18, no. 2, pp. 153-171.

Bosch, J. L. & Hunink, M. G. 1996, "The relationship between descriptive and valualational quality-of-life measures in patients with intermittent claudication", *Med Decis Making.*, vol. 16, no. 3, pp. 217-225.

Boston, N. K., Boynton, P. M., & Hood, S. 2001, "An inner city GP unit versus conventional care for elderly patients: prospective comparison of health functioning, use of services and patient satisfaction", *Fam.Pract.*, vol. 18, no. 2, pp. 141-148.

Brazier, J., Roberts, J., & Deverill, M. 2002, "The estimation of a preference-based measure of health from the SF-36", *J Health Econ.*, vol. 21, no. 2, pp. 271-292.

Brazier, J., Roberts, J., Tsuchiya, A., & Busschbach, J. 2004a, "A comparison of the EQ-5D and SF-6D across seven patient groups", *Health Econ.*, vol. 13, no. 9, pp. 873-884.

Brazier, J., Usherwood, T., Harper, R., & Thomas, K. 1998, "Deriving a preference-based single index from the UK SF-36 Health Survey", *J Clin.Epidemiol.*, vol. 51, no. 11, pp. 1115-1128.

Brazier, J. E., Kolotkin, R. L., Crosby, R. D., & Williams, G. R. 2004b, "Estimating a preference-based single index for the Impact of Weight on Quality of Life-Lite (IWQOL-Lite) instrument from the SF-6D", *Value Health.*, vol. 7, no. 4, pp. 490-498.

Brazier, J. E., Walters, S. J., Nicholl, J. P., & Kohler, B. 1996, "Using the SF-36 and Euroqol on an elderly population", *Qual.Life Res.*, vol. 5, no. 2, pp. 195-204.

Brazier, J. E., Yang, Y., & Tsuchiya, A. 2007, *A review of studies mapping (or cross walking) from non-preference based measures of health to generic preference-based measures*, School of Health and Related Research (ScHARR), University of Sheffield, 08/02.

Breeze, E., Jones, D. A., Wilkinson, P., Latif, A. M., Bulpitt, C. J., & Fletcher, A. E. 2004, "Association of quality of life in old age in Britain with socioeconomic position: baseline data from a randomised controlled trial", *J.Epidemiol.Community Health.*, vol. 58, no. 8, pp. 667-673.

Brennan, D. S. & Spencer, A. J. 2006, "Mapping oral health related quality of life to generic health state values", *BMC Health Serv.Res.*, vol. 6:96., p. 96.

Breusch, T. S. & Pagan, A. R. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* 47, 1287-1294. 1979.
Ref Type: Internet Communication

Briggs, A. H. & Gray, A. M. 1999, "Handling uncertainty when performing economic evaluation of healthcare interventions", *Health Technol.Assess.*, vol. 3, no. 2, pp. 1-134.

British Geriatrics Society. Intermediate Care: Guidance for Commissioners and Providers of Health and Social Care BGS Compendium Document 4.
www.bgs.org.uk/compendium/comd4.html . 2005. [Accessed: 22 July 2008]

Britton, A., Shipley, M., Marmot, M., & Hemingway, H. 2004, "Does access to cardiac investigation and treatment contribute to social and ethnic differences in coronary heart disease? Whitehall II prospective cohort study", *BMJ.*, vol. 329, no. 7461, p. 318.

Brooks, N., Ashton, A., & Hainsworth, B. 2003, "Pilot evaluation of an intermediate care scheme", *Nurs.Stand.*, vol. 17, no. 23, pp. 33-35.

Bult, J. R., Bosch, J. L., & Hunink, M. G. 1996, "Heterogeneity in the relationship between the standard-gamble utility measure and health-status dimensions", *Med Decis Making.*, vol. 16, no. 3, pp. 226-233.

Bult, J. R., Hunink, M. G., Tsevat, J., & Weinstein, M. C. 1998, "Heterogeneity in the relationship between the time tradeoff and Short Form-36 for HIV-infected and primary care patients", *Med Care.*, vol. 36, no. 4, pp. 523-532.

Burholt, V. 2004, "The settlement patterns and residential histories of older Gujaratis, Punjabis and Sylhetis in Birmingham, England", *Ageing & Society*, vol. 24, pp. 383-409.

Busschbach, J. J., McDonnell, J., Essink-Bot, M. L., & van Hout, B. A. 1999, "Estimating parametric relationships between health description and health valuation with an application to the EuroQol EQ-5D", *J Health Econ.*, vol. 18, no. 5, pp. 551-571.

Buxton, M. J., Lacey, L. A., Feagan, B. G., Niecko, T., Miller, D. W., & Townsend, R. J. 2007, "Mapping from disease-specific measures to utility: an analysis of the relationships between the Inflammatory Bowel Disease Questionnaire and Crohn's Disease Activity Index in Crohn's disease and measures of utility", *Value Health.*, vol. 10, no. 3, pp. 214-220.

Campbell, S. E., Seymour, D. G., & Primrose, W. R. 2004, "A systematic literature review of factors affecting outcome in older medical patients admitted to hospital", *Age Ageing.*, vol. 33, no. 2, pp. 110-115.

Cantoni, E. & Ronchetti, E. 2006, "A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures", *J Health Econ.*, vol. 25, no. 2, pp. 198-213.

Carpenter, I., Kotiadis, K., Kotiadis, K., & Mackenzie, M. 2003, *An Evaluation of Intermediate Care Services for Older People*, Centre for Health Services Studies, University of Kent, Kent.

Casado-Diaz, M. A., Kaiser, C., & Warnes, A. M. 2004, "Northern European retired residents in nine southern European areas: characteristics, motivations and adjustment", *Ageing & Society*, vol. 24, pp. 353-381.

Chancellor, J. V., Coyle, D., & Drummond, M. F. 1997, "Constructing health state preference values from descriptive quality of life outcomes: mission impossible?", *Qual Life Res.*, vol. 6, no. 2, pp. 159-168.

Chay, K. Y. & Powell, J. L. 2001, "Semiparametric censored regression models", *Journal of Economic Perspectives*, vol. 15, pp. 29-42.

Chen, J. & Shao, J. 2000, "Nearest neighbor imputation for survey data", *J Off Stat*, vol. 16, no. 2, pp. 113-131.

Chen, J. S., March, L. M., Schwarz, J., Zochling, J., Makaroff, J., Sitoh, Y. Y., Lau, T. C., Lord, S. R., Cameron, I. D., Cumming, R. G., & Sambrook, P. N. 2005, "A multivariate regression model predicted falls in residents living in intermediate hostel care", *Journal of Clinical Epidemiology*, vol. 58, no. 5, pp. 503-508.

Cheung, Y. B., Tan, L. C., Lau, P. N., Au, W. L., & Luo, N. 2008, "Mapping the eight-item Parkinson's Disease Questionnaire (PDQ-8) to the EQ-5D utility index", *Qual Life Res.*, vol. 17, no. 9, pp. 1173-1181.

Cheung, Y. B., Thumboo, J., Gao, F., Ng, G. Y., Pang, G., Koo, W. H., Sethi, V. K., Wee, J., & Goh, C. 2009, "Mapping the English and Chinese Versions of the

- Functional Assessment of Cancer Therapy-General to the EQ-5D Utility Index", *Value Health.*, vol. 12, no. 2, pp.371-376
- Clarke, P., Gray, A., & Holman, R. 2002, "Estimating utility values for health states of type 2 diabetic patients using the EQ-5D (UKPDS 62)", *Med Decis Making.*, vol. 22, no. 4, pp. 340-349.
- Clayson, D. J., Briggs, A. H., Sculpher, M., & De Hert, M. 2004, "Mapping utility scores from the EQ-5D and SF-6D onto the schizophrenia quality of life scale", *Value Health.*, vol. 7, p. 277.
- Clough, R., Manthorpe, J., Green, B., Fox, D., Raymond, G., Wilson, P., Raymond, V., Sumner, K., Bright, L., & Hay, J. 2007, *The support older people want and the services they need*, Joseph Rowntree Foundation, York.
- Coast, J., Peters, T. J., Richards, S. H., & Gunnell, D. J. 1998a, "Use of the EuroQoL among elderly acute care patients", *Qual.Life Res.*, vol. 7, no. 1, pp. 1-10.
- Coast, J., Richards, S. H., Peters, T. J., Gunnell, D. J., Darlow, M. A., & Pounsford, J. 1998b, "Hospital at home or acute hospital care? A cost minimisation analysis", *BMJ.*, vol. 316, no. 7147, pp. 1802-1806.
- Cohen, J. & Cohen, P. 1983, *Applied multiple regression/correlation analysis for the behavioral sciences* , 2nd edn, Erlbaum, Hillsdale, NJ.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. 2003, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd edition. 3rd edn, Lawrence Erlbaum, Mahwah, N.J.
- Collins, L. M., Schafer, J. L., & Kam, C. M. 2001, "A comparison of inclusive and restrictive strategies in modern missing-data procedures 270", *Psychological Methods*, vol. 6, pp. 330-351.
- Cook, R. J., Zeng, L., & Yi, G. Y. 2004, "Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation", *Biometrics.*, vol. 60, no. 3, pp. 820-828.
- Cornes, M. & Clough, R. 2001, "The Continuum of Care: older people's experiences of intermediate care.", *Education and Ageing*, vol. 16, no. 2.
- Cotton, M. M., Bucknall, C. E., Dagg, K. D., Johnson, M. K., MacGregor, G., Stewart, C., & Stevenson, R. D. 2000, "Early discharge for patients with exacerbations of chronic obstructive pulmonary disease: a randomized controlled trial", *Thorax.*, vol. 55, no. 11, pp. 902-906.
- Cowpe, J. 2005, "Intermediate Care: Policy and Context," in *Intermediate and Continuing Care: Policy and Practice*, B. Roe & R. Beech, eds., Blackwell Publishing Ltd, Oxford.
- Croninger, R. G. & Douglas, K. M. 2005, "Missing Data and Institutional Research ," in *Survey research. Emerging issues. New directions for institutional research #127.*, P. D. Umbach, ed., Jossey-Bass, San Fransisco, pp. 33-50.

- Curran, D., Bacchi, M., Schmitz, S. F., Molenberghs, G., & Sylvester, R. J. 1998, "Identifying the types of missingness in quality of life data from clinical trials", *Stat.Med.*, vol. 17, no. 5-7, pp. 739-756.
- Curtis, L. & Netten, A. Unit Costs of Health Social Care . 2004. Canterbury, PSSRU, University of Kent.
- D'Agostino, R. B. J. & Rubin, D. B. 2008, "Estimating and using propensity scores with partially missing data.", *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 749-759.
- Daniel, W. W. & Terrell, J. C. 1995, *Business Statistics*, Seventh edn, Houghton Mifflin Company, Boston.
- David, M., Little, R. J. A., Samuhel, M. E., & Triest, R. K. 1986, "Alternative Methods for CPS Income Imputation", *Journal of the American Statistical Association*, vol. 81, pp. 29-41.
- Davies, L., Wilkinson, M., Bonner, S., Calverley, P. M., & Angus, R. M. 2000, ""Hospital at home" versus hospital care in patients with exacerbations of chronic obstructive pulmonary disease: prospective randomised controlled trial", *BMJ.*, vol. 321, no. 7271, pp. 1265-1268.
- Davies, P. 1996, "Social approaches to health outcomes," in *Health Outcomes: Biology, Social, and Economic Perspectives*, H. M. Macbeth, ed., Oxford University Press, New York, NY, pp. 100-105.
- Department of Health 1999, *Royal Commission For Elderly People*, The Stationery Office, London.
- Department of Health 2000a, *Shaping the Future NHS: Long Term Planning for Hospital and Related Services. Consultation Document on the Findings of the National Beds Inquiry.*, The Stationery Office, London.
- Department of Health 2000b, *Shaping the Future NHS: Long-term Planning for Hospitals and Related Services. National Beds Inquiry.* Department of Health, London.
- Department of Health 2000c, *The NHS Plan. A Plan for Investment. A Plan for Reform*, The Stationery Office, London.
- Department of Health 2001a, *Intermediate Care Services for Older People, HSC 2001/01.*, The Stationery Office, London.
- Department of Health 2001b, *National Service Framework for Older People*, Department of Health, London.
- Department of Health 2002a, "Delivering the NHS Plan: Next Steps on Investment, Next Steps on Reform.," Department of Health, London.
- Department of Health 2002b, *Intermediate Care: Moving Forward*, The Stationery Office, London.

Department of Health 2002c, *National Service Framework for Older People - Supporting Implementation. Intermediate Care: Moving Forward*. Department of Health., London.

Department of Health 2002d, *Shaping the Future NHS: Long-term Planning for Hospitals and Related Services. National Beds Inquiry.*, Department of Health, London.

Department of Health 2003, "Discharge from hospital: Pathway, Process and Practice." The Stationery Office, London.

Department of Health 2005, *The 2004/05 National Survey of Investment in Mental Health Services*, Department of Health, London.

Diehr, P., Johnson, L. L., Patrick, D. L., & Psaty, B. 2005, "Methods for incorporating death into health-related variables in longitudinal studies", *J.Clin.Epidemiol.*, vol. 58, no. 11, pp. 1115-1124.

Diggle, P. & Kenwood, M. G. 1994, "Informative dropout in longitudinal data analysis (with discussion). ", *Applied Statistics*, vol. 43, pp. 49-94.

Dixon, S., McEwan, P., & Currie, C. 2003, "Estimating the health utility of treatment in adults with growth hormone deficiency", *Journal of Drug Assessment*, vol. 6, no. 1, pp. 1-12.

Dobrez, D., Cella, D., Pickard, A. S., Lai, J. S., & Nickolov, A. 2007, "Estimation of patient preference-based utility weights from the functional assessment of cancer therapy - general", *Value Health.*, vol. 10, no. 4, pp. 266-272.

Dolan, P., Gudex, C., Kind, P., & Williams, A. 1996, "The time trade-off method: results from a general population study", *Health Econ.*, vol. 5, no. 2, pp. 141-154.

Dolan, P. 1997, "Modeling valuations for EuroQol health states", *Med Care.*, vol. 35, no. 11, pp. 1095-1108.

Downing, A. & Wilson, R. 2005, "Older people's use of Accident and Emergency services", *Age and Ageing*, vol. 34, no. 1, pp. 24-30.

Draper, N. R. & Smith, H. 1998, *Applied Regression Analysis*, 3rd edn, John Wiley & Sons, New York, NY.

Drever, F., Doran, T., & Whitehead, M. 2004, "Exploring the relation between class, gender, and self rated general health using the new socioeconomic classification. A study using data from the 2001 census", *J Epidemiol.Community Health.*, vol. 58, no. 7, pp. 590-596.

Drummond, M., O'Brien, B., Stoddart, G., & Torrance, G. 1997, *Methods for the economic evaluation of health care programmes*, second edn, Oxford University Press, Oxford.

- Duan, N. 1983, "Smearing estimate a nonparametric retransformation method", *J.Amer.Statist.Assoc.*, vol. 78, pp. 605-610.
- Durrant, G. B. 2005, *Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review*, ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute, University of Southampton.
- Ebrahim, S., Papacosta, O., Wannamethee, G., & Adamson, J. 2004, "Social inequalities and disability in older men: prospective findings from the British regional heart study", *Soc.Sci.Med.*, vol. 59, no. 10, pp. 2109-2120.
- Elkan, R., Kendrick, D., Dewey, M., Hewitt, M., Robinson, J., Blair, M., Williams, D., & Brummell, K. 2001, "Effectiveness of home based support for older people: systematic review and meta-analysis", *British Medical Journal*, vol. 323, no. 7315, pp. 719-724B.
- Ellis, A., Trappes-Lomax, T., Fox, M., Taylor, R., Power, M., Stead, J., & Bainbridge, I. 2006, "Buying time II: an economic evaluation of a joint NHS/Social Services residential rehabilitation unit for older people on discharge from hospital", *Health Soc.Care Community.*, vol. 14, no. 2, pp. 95-106.
- European Institute of Women's Health. Demographic Trends.
<http://www.eurohealth.ie/newrep/demograph.htm> . 1996. [Accessed 28 July 2008]
- Evandrou, M. 2000, "Social inequalities in later life: the socio-economic position of older people from ethnic minority groups in Britain", *Popul.Trends.* no. 101, pp. 11-18.
- Evans, O., Singleton, N., & Melzer, H. 2003, *The mental health of older people - Report based on the analysis of the ONS Survey of Psychiatric Morbidity among Adults in Great Britain carried out in 2000*, Office for National Statistics, London.
- Fayers, P. M., Curran, D., & Machin, D. 1998, "Incomplete quality of life data in randomized trials: missing items", *Stat.Med.*, vol. 17, no. 5-7, pp. 679-696.
- Feeny, D. H. & Torrance, G. W. 1989, "Incorporating utility-based quality-of-life assessment analysis.", *Medical Care*, vol. 27, p. S190-S204.
- Fleming, S. A., Blake, H., Gladman, J. R., Hart, E., Lymbery, M., Dewey, M. E., McCloughry, H., Walker, M., & Miller, P. 2004, "A randomised controlled trial of a care home rehabilitation service to reduce long-term institutionalisation for elderly people", *Age Ageing.*, vol. 33, no. 4, pp. 384-390.
- Fletcher, A. E., Price, G. M., Ng, E. S., Stirling, S. L., Bulpitt, C. J., Breeze, E., Nunes, M., Jones, D. A., Latif, A., Fasey, N. M., Vickers, M. R., & Tulloch, A. J. 2004, "Population-based multidimensional assessment of older people in UK general practice: a cluster-randomised factorial trial", *Lancet.*, vol. 364, no. 9446, pp. 1667-1677.

- Forder, J. & Netten, A. 2000, "The price of placements in residential and nursing home care: the effects of contracts and competition", *Health Econ.*, vol. 9, no. 7, pp. 643-657.
- Forster, A., Young, J., Lambley, R., & Langhorne, P. 2008, "Medical day hospital care for the elderly versus alternative forms of care", *Cochrane Database Syst.Rev.* no. 4, p. CD001730.
- Forster, J. J. & Smith, P. W. F. 1998, "Model-based inference for categorical survey data subject to non-ignorable non-response", *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, vol. 60, pp. 57-60.
- Foster, E. M. & Bickman, L. 1996, "An Evaluator's Guide To Detecting Attrition Problems ", *Evaluation Review*, vol. 20, no. 6, pp. 695-723.
- Foster, E. M. & Fang, G. Y. 2004, "Alternative methods for handling attrition: an illustration using data from the Fast Track evaluation", *Eval.Rev.*, vol. 28, no. 5, pp. 434-464.
- Fox-Wasylyshyn, S. M. & El-Masri, M. M. 2005, "Handling missing data in self-report measures", *Res.Nurs.Health.*, vol. 28, no. 6, pp. 488-495.
- Franks, P., Lubetkin, E. I., Gold, M. R., & Tancredi, D. J. 2003, "Mapping the SF-12 to preference-based instruments: convergent validity in a low-income, minority population", *Med Care.*, vol. 41, no. 11, pp. 1277-1283.
- Franks, P., Lubetkin, E. I., Gold, M. R., Tancredi, D. J., & Jia, H. 2004, "Mapping the SF-12 to the EuroQol EQ-5D Index in a national US sample", *Med Decis.Making.*, vol. 24, no. 3, pp. 247-254.
- Froberg, D. G. & Kane, R. L. 1989, "Methodology for measuring health-state preferences--I: Measurement strategies", *J Clin.Epidemiol.*, vol. 42, no. 4, pp. 345-354.
- Fryback, D. G., Lawrence, W. F., Martin, P. A., Klein, R., & Klein, B. E. 1997, "Predicting Quality of Well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study", *Med Decis Making.*, vol. 17, no. 1, pp. 1-9.
- Gauggel, S., Lammler, G., Borchelt, M., Steinhagen-Thiessen, E., Bocker, M., & Heinemann, A. 2002, "[Agreement on the Barthel Index. A rapid analysis of other and self-assessment in elderly stroke patients]", *Z.Gerontol.Geriatr.*, vol. 35, no. 2, pp. 102-110.
- Gilbertson, L., Langhorne, P., Walker, A., Allen, A., & Murray, G. D. 2000, "Domiciliary occupational therapy for patients with stroke discharged from hospital: randomised controlled trial", *BMJ.*, vol. 320, no. 7235, pp. 603-606.
- Glynn, R. J., Laird, N. M., & Rubin, D. B. 1986, "Drawing Inferences from Self-selected Samples," in *Selection modelling versus mixture modelling with nonignorable nonresponse*, H. Wainer, ed., Springer-Verlag, New York, pp. 115-142.

- Godfrey, M. & Townsend, J. 2008, "Older people in transition from illness to health: trajectories of recovery", *Qual.Health Res.*, vol. 18, no. 7, pp. 939-951.
- Gold, M. 1996, "Panel on cost-effectiveness in health and medicine", *Med Care.*, vol. 34, no. 12 Suppl, p. DS197-DS199.
- Gosman-Hedstrom, G. & Svensson, E. 2000, "Parallel reliability of the functional independence measure and the Barthel ADL index", *Disabil.Rehabil.*, vol. 22, no. 16, pp. 702-715.
- Gray, A. M., Rivero-Arias, O., & Clarke, P. M. 2006, "Estimating the association between SF-12 responses and EQ-5D utility values by response mapping", *Med Decis Making.*, vol. 26, no. 1, pp. 18-29.
- Green, J., Young, J., Forster, A., Mallinder, K., Bogle, S., Lowson, K., & Small, N. 2005, "Effects of locality based community hospital care on independence in older people needing rehabilitation: randomised controlled trial", *BMJ.*, vol. 331, no. 7512, pp. 317-322.
- Greene, W. 1997, *Econometric Analysis*, Third edn, Prentice-Hall International, Inc, New Jersey.
- Greenless, J. S., Reece, W. S., & Zieschang, K. D. 1982, "Imputation of Missing Values When the Probability of Response Depends On the Value Being Imputed. , 269", *Journal of the American Statistical Association*, vol. 77, pp. 251-261.
- Griffiths, P., Edwards, M., Forbes, A., & Harris, R. 2005, "Post-acute intermediate care in nursing-led units: a systematic review of effectiveness", *Int.J Nurs.Stud.*, vol. 42, no. 1, pp. 107-116.
- Griffiths, P. & Evans, A. 1995, *Evaluating a Nursing Led Inpatient Service: an Interim Report*, King's Fund Centre, London.
- Griffiths, P., Harris, R., Richardson, G., Hallett, N., Heard, S., & Wilson-Barnett, J. 2001, "Substitution of a nursing-led inpatient unit for acute services: randomized controlled trial of outcomes and cost of nursing-led intermediate care", *Age Ageing.*, vol. 30, no. 6, pp. 483-488.
- Griffiths, P. & Wilson-Barnett, J. 1998, "The effectiveness of 'nursing beds': a review of the literature", *J.Adv.Nurs.*, vol. 27, no. 6, pp. 1184-1192.
- Griffiths, P. & Wilson-Barnett, J. 2000, "Influences on length of stay in intermediate care: lessons from the nursing-led inpatient unit studies", *Int.J.Nurs.Stud.*, vol. 37, no. 3, pp. 245-255.
- Griffiths, P., Wilson-Barnett, J., Richardson, G., Spilsbury, K., Miller, F., & Harris, R. 2000, "The effectiveness of intermediate care in a nursing-led in-patient unit", *Int.J Nurs.Stud.*, vol. 37, no. 2, pp. 153-161.
- Griffiths, P. D., Edwards, M. H., Forbes, A., Harris, R. L., & Ritchie, G. 2007, "Effectiveness of intermediate care in nursing-led in-patient units", *Cochrane Database of Systematic Reviews* no. 2.

- Grimley, E. J. & Tallis, R. C. 2001, "A new beginning for care for elderly people?", *BMJ.*, vol. 322, no. 7290, pp. 807-808.
- Grootendorst, P., Marshall, D., Pericak, D., Bellamy, N., Feeny, D., & Torrance, G. W. 2007, "A model to estimate health utilities index mark 3 utility scores from WOMAC index scores in patients with osteoarthritis of the knee", *J Rheumatol.*, vol. 34, no. 3, pp. 534-542.
- Grundy, E. & Sloggett, A. 2003, "Health inequalities in the older population: the role of personal capital, social resources and socio-economic circumstances", *Soc.Sci.Med.*, vol. 56, no. 5, pp. 935-947.
- Gujarati, D. 1995, *Basic Econometrics*, Third edn, McGraw-Hill, Inv, New York.
- Gunnell, D., Coast, J., Richards, S. H., Peters, T. J., Pounsford, J. C., & Darlow, M. A. 2000, "How great a burden does early discharge to hospital-at-home impose on carers? A randomized controlled trial", *Age Ageing.*, vol. 29, no. 2, pp. 137-142.
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. 1998, *Multivariate data analysis with readings*, 4th edn, Prentice Hall, Upper Saddle River, NJ.
- Hardy, R. & Kuh, D. 2005, "Social and environmental conditions across the life course and age at menopause in a British birth cohort study", *BJOG.*, vol. 112, no. 3, pp. 346-354.
- Harris, R., Richardson, G., Griffiths, P., Hallett, N., & Wilson-Barnett, J. 2005, "Economic evaluation of a nursing-led inpatient unit: the impact of findings on management decisions of service utility and sustainability", *J Nurs.Manag.*, vol. 13, no. 5, pp. 428-438.
- Harris, R., Wilson-Barnett, J., & Griffiths, P. 2007, "Effectiveness of nursing-led inpatient care for patients with post-acute health care needs: secondary data analysis from a programme of randomized controlled trials", *J Eval.Clin.Pract.*, vol. 13, no. 2, pp. 198-205.
- Healey, F., Monro, A., Cockram, A., Adams, V., & Heseltine, D. 2004, "Using targeted risk factor reduction to prevent falls in older in-patients: a randomised controlled trial", *Age Ageing.*, vol. 33, no. 4, pp. 390-395.
- Heckman, J. J. 1976, "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models", *Annals of Economic and Social Measurement*, vol. 5, pp. 475-492.
- Heckman, J. J. 1979, "Sample selection bias as a specification error", *Econometrica*, vol. 47, pp. 153-161.
- Hedeker, D. & Gibbons, R. D. 1997, "Application of random-effects pattern-mixture models for missing data in longitudinal studies", *Psychological Methods*, vol. 2, no. 1, pp. 64-78.
- Heitjan, D. F. 1997, "Annotation: what can be done about missing data? Approaches to imputation", *Am J Public Health.*, vol. 87, no. 4, pp. 548-550.

- Heyting, A., Tolboom, J. T. B. M., & Essers, J. G. A. 2008, "Statistical Handling of Drop-Outs in Longitudinal Clinical Trials", *Statistics in Medicine*, vol. 11, pp. 2043-2061.
- Hilari, K., Owen, S., & Farrelly, S. J. 2007, "Proxy and self-report agreement on the Stroke and Aphasia Quality of Life Scale-39", *J Neurol.Neurosurg.Psychiatry.*, vol. 78, no. 10, pp. 1072-1075.
- HM Treasury 2002, *Securing our Future Health: Taking a Long-term View. Final Report (The Wanless Report).*, HM Treasury, London.
- Hogan, J. W. & Laird, N. M. 1997, "Mixture models for the joint distribution of repeated measures and event times", *Stat Med.*, vol. 16, no. 1-3, pp. 239-257.
- Hollingworth, W., Deyo, R. A., Sullivan, S. D., Emerson, S. S., Gray, D. T., & Jarvik, J. G. 2002, "The practicality and validity of directly elicited and SF-36 derived health state preferences in patients with low back pain", *Health Econ.*, vol. 11, no. 1, pp. 71-85.
- Hollis, S. 2002, "A graphical sensitivity analysis for clinical trials with non-ignorable missing binary outcome", *Stat.Med.*, vol. 21, pp. 3823-3834.
- Huang, I. C., Frangakis, C., Atkinson, M. J., Willke, R. J., Leite, W. L., Vogel, W. B., & Wu, A. W. 2008, "Addressing ceiling effects in health status measures: a comparison of techniques applied to measures for people with HIV disease", *Health Serv.Res.*, vol. 43, no. 1 Pt 1, pp. 327-339.
- Hung, S. Y., Pickard, A. S., Witt, W. P., & Lambert, B. L. 2007, "Pain and depression in caregivers affected their perception of pain in stroke patients", *J Clin Epidemiol.*, vol. 60, no. 9, pp. 963-970.
- Hyde, C. J., Robert, I. E., & Sinclair, A. J. 2000, "The effects of supporting discharge from hospital to home in older people", *Age Ageing.*, vol. 29, no. 3, pp. 271-279.
- Hyndman, R. J. & Koehler, A. B. 2006, "Another look at measures of forecast accuracy", *International Journal of Forecasting*, vol. 22, no. 4, pp. 679-688.
- ICNET 2005, *A National Evaluation of the Costs and Outcomes of Intermediate Care for Older People: Final Report*, The University of Leicester, Leicester.
- Institute of Health Sciences and Public Health Research 2005, *An Evaluation of Intermediate Care for Older People, Final Report*, University of Leeds, Leeds.
- Jenkinson, D. & Ford, G. A. 2006, "Research and development in stroke services 130", *BMJ.*, vol. 332, no. 7537, p. 318.
- Johnson, J. A. & Coons, S. J. 1998, "Comparison of the EQ-5D and SF-12 in an adult US sample", *Qual.Life Res.*, vol. 7, no. 2, pp. 155-166.
- Johnston, J. 1997, *Econometric Methods*. McGraw-Hill, New York.

- Jones, A. M. 2000, "Health Econometrics," in *Handbook of Health Economics*, J. P. Newhouse & A. J. Culyer, eds., Elsevier, Amsterdam.
- Kaambwa, B., Bryan, S., Barton, P., Parker, H., Martin, G., Hewitt, G., Parker, S., & Wilson, A. 2008, "Costs and health outcomes of intermediate care: results from five UK case study sites", *Health Soc.Care Community.*, vol. 16, no. 6, pp. 573-581.
- Kaplan, R. M. & Bush, J. W. 1982, "Health-related quality of life measurement for evaluation research and policy analysis.", *Health Psychology* pp. 61-80.
- Kaplan, R. M., Groessl, E. J., Sengupta, N., Sieber, W. J., & Ganiats, T. G. 2005, "Comparison of measured utility scores and imputed scores from the SF-36 in patients with rheumatoid arthritis", *Med Care.*, vol. 43, no. 1, pp. 79-87.
- Kendall, M. G. & Buckland, W. R. 1971, *A Dictionary Of Statistical Terms* 178 Hafner Publishing Company, New York.
- Kennedy, C. E. & Aoki, N. 2002, "Generating a mortality model from a pediatric ICU (PICU) database utilizing knowledge discovery", *Proc.AMIA.Symp.* pp. 375-379.
- Kilian, R., Matschinger, H., Loeffler, W., Roick, C., & Angermeyer, M. C. 2002, "A comparison of methods to handle skew distributed cost variables in the analysis of the resource consumption in schizophrenia treatment", *J Ment.Health Policy Econ.*, vol. 5, no. 1, pp. 21-31.
- Kind, P., Hardman, G., & Macran, S. 1999, *UK population norms for EQ-5D. Discussion paper 172*, Centre for Health Economics, University of York.
- Kline, R. 1998, *Principles and practice of structural equation modeling* The Guilford Press, New York.
- Kmetz, A., Joseph, L., Berger, C., & Tenenhouse, A. 2002, "Multiple imputation to account for missing data in a survey: estimating the prevalence of osteoporosis", *Epidemiology.*, vol. 13, no. 4, pp. 437-444.
- Koltowska-Haggstrom, M., Jonsson, B., Isacson, D., & Bingefors, K. 2007, "Using EQ-5D to derive general population-based utilities for the quality of life assessment of growth hormone deficiency in adults (QoL-AGHDA)", *Value Health.*, vol. 10, no. 1, pp. 73-81.
- Kulkarni, A. V. 2006, "Distribution-based and anchor-based approaches provided different interpretability estimates for the Hydrocephalus Outcome Questionnaire", *J Clin.Epidemiol.*, vol. 59, no. 2, pp. 176-184.
- Kwon, S., Hartzema, A. G., Duncan, P. W., & Min-Lai, S. 2004, "Disability measures in stroke: relationship among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale", *Stroke.*, vol. 35, no. 4, pp. 918-923.
- Laird, N. M. 1988, "Missing data in longitudinal studies", *Stat.Med.*, vol. 7, no. 1-2, pp. 305-315.

Lambert, M. & Arblaster, L. 2000, *Factors associated with acute use of hospital beds by older people: a systematic review of the literature. Shaping the future NHS: long-term planning for hospitals and related services.* Department of Health, London, Consultation on the Findings of the National Beds Inquiry-Supporting Analysis.

Lawrence, W. F. & Fleishman, J. A. 2004, "Predicting EuroQoL EQ-5D preference scores from the SF-12 Health Survey in a nationally representative sample", *Med Decis.Making.*, vol. 24, no. 2, pp. 160-169.

Lenert, L. A., Rupnow, M. F., & Elnitsky, C. 2005, "Application of a disease-specific mapping function to estimate utility gains with effective treatment of schizophrenia", *Health Qual Life Outcomes.*, vol. 3:57., p. 57.

Lenert, L. A., Sturley, A. P., Rapaport, M. H., Chavez, S., Mohr, P. E., & Rupnow, M. 2004, "Public preferences for health states with schizophrenia and a mapping function to estimate utilities from positive and negative symptom scale scores", *Schizophr.Res.*, vol. 71, no. 1, pp. 155-165.

Leon, A. C., Mallinckrodt, C. H., Chuang-Stein, C., Archibald, D. G., Archer, G. E., & Chartier, K. 2006, "Attrition in randomized controlled clinical trials: methodological issues in psychopharmacology", *Biol.Psychiatry.*, vol. 59, no. 11, pp. 1001-1005.

Little, R. J. A. 1992, "Regression with missing X's: a review 158", *Journal of the American Statistical Association*, vol. 87, pp. 1227-1237.

Little, R. J. A. 1995, "Modelling the drop-out mechanism in repeated measures studies", *Journal of the American Statistical Association*, vol. 90, pp. 1112-1121.

Little, R. J. A. & Rubin, D. B. 1987, *Statistical Analysis with Missing Data* John Wiley., New York.

Little, R. J. A. & Rubin, D. B. 2002, *Statistical Analysis with Missing Data* John Wiley, New York.

Little, R. J. A. & Schenker, N. 1995, "Missing Data," in *Handbook of Statistical Modeling in the Social and Behavioral Sciences*, G. Arminger, C. C. Clogg, & M. E. Sobel, eds., Plenum, New York, pp. 39-75.

Lobo, F. S., Gross, C. R., & Matthees, B. J. 2004, "Estimation and comparison of derived preference scores from the SF-36 in lung transplant patients", *Qual Life Res.*, vol. 13, no. 2, pp. 377-388.

Long, S. 1997, *Regression Models for Categorical and Limited Dependent Variables. Advanced Quantitative Techniques in the Social Sciences* 7 Sage Publications, London.

Long, K., Sudha, S., & Mutran, E. J. 1998, "Elder-proxy agreement concerning the functional status and medical history of the older person: the impact of caregiver burden and depressive symptomatology", *J Am Geriatr Soc.*, vol. 46, no. 9, pp. 1103-1111.

- Longworth, L., Buxton, M. J., Sculpher, M., & Smith, D. H. 2005, "Estimating utility data from clinical indicators for patients with stable angina", *Eur.J Health Econ.*, vol. 6, no. 4, pp. 347-353.
- Lundberg, L., Johannesson, M., Isacson, D. G., & Borgquist, L. 1999, "The relationship between health-state utilities and the SF-12 in a general population", *Med Decis.Making.*, vol. 19, no. 2, pp. 128-140.
- Lyons, R. A., Crome, P., Monaghan, S., Killalea, D., & Daley, J. A. 1997, "Health status and disability among elderly people in three UK districts", *Age Ageing.*, vol. 26, no. 3, pp. 203-209.
- Mahoney, F. I. & Barthel, D. W. 1965, "Functional Evaluation: The Barthel Index", *Md State Med J.*, vol. 14:61-5., pp. 61-65.
- Mallinckrodt, C. H., Sanger, T. M., Dube, S., DeBrot, D. J., Molenberghs, G., Carroll, R. J., Potter, W. Z., & Tollefson, G. D. 2003, "Assessing and interpreting treatment effects in longitudinal clinical trials with missing data", *Biol.Psychiatry.*, vol. 53, no. 8, pp. 754-760.
- Manca, A. & Palmer, S. 2005, "Handling missing data in patient-level cost-effectiveness analysis alongside randomised clinical trials 155", *Appl.Health Econ.Health Policy.*, vol. 4, no. 2, pp. 65-75.
- Mann, R., Birks, Y., Hall, J., Torgerson, D., & Watt, I. 2006, "Exploring the relationship between fear of falling and neuroticism: a cross-sectional study in community-dwelling women over 70", *Age Ageing.*, vol. 35, no. 2, pp. 143-147.
- Manning, W. G. 1998, "The logged dependent variable, heteroscedasticity, and the retransformation problem", *J Health Econ.*, vol. 17, no. 3, pp. 283-295.
- Manning, W. G. & Mullahy, J. 2001b, "Estimating log models: to transform or not to transform?", *J.Health Econ.*, vol. 20, no. 4, pp. 461-494.
- Marshall, D., Pericak, D., Grootendorst, P., Gooch, K., Faris, P., Frank, C., Bellamy, N., Torrance, G., & Feeny, D. 2008, "Validation of a prediction model to estimate health utilities index Mark 3 utility scores from WOMAC index scores in patients with osteoarthritis of the hip", *Value Health.*, vol. 11, no. 3, pp. 470-477.
- Matthews, F. & Brayne, C. 2005, "The incidence of dementia in England and Wales: Findings from the five identical sites of the MRC CFA study", *Plos Medicine*, vol. 2, no. 8, pp. 753-763.
- Matthews, F. E., Chatfield, M., Freeman, C., McCracken, C., & Brayne, C. 2004, "Attrition and bias in the MRC cognitive function and ageing study: an epidemiological investigation", *BMC Public Health.*, vol. 4:12., p. 12.
- Matthews, R. J., Smith, L. K., Hancock, R. M., Jagger, C., & Spiers, N. A. 2005, "Socioeconomic factors associated with the onset of disability in older age: a longitudinal study of people aged 75 years and over", *Soc.Sci.Med.*, vol. 61, no. 7, pp. 1567-1575.

- Mayhew, L. & Lawrence, D. 2006, "The costs and service implications of substituting intermediate care for acute hospital care", *Health Serv Manage.Res.*, vol. 19, no. 2, pp. 80-93.
- McCullagh, P. & Nelder, J. A. 1989, *Generalized linear models*, 2nd edn, Chapman & Hall, London.
- McDonough, C. M., Grove, M. R., Tosteson, T. D., Lurie, J. D., Hilibrand, A. S., & Tosteson, A. N. 2005, "Comparison of EQ-5D, HUI, and SF-36-derived societal health state values among spine patient outcomes research trial (SPORT) participants", *Qual Life Res.*, vol. 14, no. 5, pp. 1321-1332.
- McEwan, P., Kind, P., Dixon, S., & Currie, S.J. 2004, "Using an artificial neural network to predict utility scores from SF36 data", *Value Health.*, vol. 7, no. 3, p. 305.
- McEwan, P., Morrissey, M., & Currie, C. J. 2003, "An improved model to calculate utility scores from SF36 data", *Value Health.*, vol. 6, pp. 302-303.
- Melis, R. J., Olde Rikkert, M. G., Parker, S. G., & van Eijken, M. I. 2004, "What is intermediate care?", *BMJ.*, vol. 329, no. 7462, pp. 360-361.
- Melsop, K. A., Boothroyd, D. B., & Hlatky, M. A. 2003, "Quality of life and time trade-off utility measures in patients with coronary artery disease", *Am Heart J.*, vol. 145, no. 1, pp. 36-41.
- Miller, L. & Singer, M. 2004, "Relationship of quality of life domains to preference-based measures of quality of life", *Value Health.*, vol. 7, p. 304.
- Miller, P., Gladman, J. R., Cunliffe, A. L., Husbands, S. L., Dewey, M. E., & Harwood, R. H. 2005, "Economic analysis of an early discharge rehabilitation service for older people", *Age Ageing.*, vol. 34, no. 3, pp. 274-280.
- Miller, R. B. & Wright, D. W. 1995, "Detecting and correcting attrition bias in longitudinal family research", *Journal of Marriage and the Family*, vol. 57, no. 4, pp. 921-929.
- Mold, F., McKevitt, C., & Wolfe, C. 2003, "A review and commentary of the social factors which influence stroke care: issues of inequality in qualitative literature", *Health Soc.Care Community.*, vol. 11, no. 5, pp. 405-414.
- Molenberghs, G., Goetghebeur, E. J., Lipsitz, S. R., Kenward, M. G., Lesaffre, E., & Michiels, B. 1999, "Missing data perspectives of the fluvoxamine data set: a review", *Stat.Med.*, vol. 18, no. 17-18, pp. 2449-2464.
- Mortimer, D. & Segal, L. 2008, "Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into QALY-weights", *Med Decis.Making.*, vol. 28, no. 1, pp. 66-89.
- Mortimer, D., Segal, L., Hawthorne, G., & Harris, A. 2007, "Item-based versus subscale-based mappings from the SF-36 to a preference-based quality of life measure", *Value Health.*, vol. 10, no. 5, pp. 398-407.

- Mugford, M. 2001, "Using systematic reviews for economic evaluation," in *Systematic reviews in healthcare: meta-analysis in context*, M. Egger, G. Davey Smith, & D. G. Altman, eds., BMJ Books, London, pp. 419-428.
- Mullahy, J. 1997, "Heterogeneity, Excess Zeros, and the Structure of Count Data Models", *Journal of Applied Econometrics*, vol. 12, pp. 337-350.
- Mullahy, J. 1998, "Much ado about two: reconsidering retransformation and the two-part model in health econometrics", *J Health Econ.*, vol. 17, no. 3, pp. 247-281.
- Mulrow, C. D., Cook, D. J., & Davidoff, F. 1997, "Systematic reviews: critical links in the great chain of evidence", *Ann.Intern.Med.*, vol. 126, no. 5, pp. 389-391.
- Murphy, R., Sackley, C. M., Miller, P., & Harwood, R. H. 2001, "Effect of experience of severe stroke on subjective valuations of quality of life after stroke", *J Neurol.Neurosurg.Psychiatry.*, vol. 70, no. 5, pp. 679-681.
- Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. 2002, "A comparison of imputation techniques for handling missing data", *West J Nurs.Res.*, vol. 24, no. 7, pp. 815-829.
- Muus, I., Petzold, M., & Ringsberg, K. C. 2009, "Health-related quality of life after stroke: reliability of proxy responses", *Clin Nurs Res.*, vol. 18, no. 2, pp. 103-118.
- Newton, J. L., Kyle, P., Liversidge, P., Robinson, G., Wilton, K., & Reeve, P. 2006, "The costs of falls in the community to the North East Ambulance Service 163", *Emerg.Med.J.*, vol. 23, no. 6, pp. 479-481.
- Nichol, M. B., Sengupta, N., & Globe, D. R. 2001, "Evaluating quality-adjusted life years: estimation of the health utility index (HUI2) from the SF-36", *Med Decis Making.*, vol. 21, no. 2, pp. 105-112.
- Norman, P., Boyle, P., & Rees, P. 2005, "Selective migration, health and deprivation: a longitudinal analysis", *Soc.Sci.Med.*, vol. 60, no. 12, pp. 2755-2771.
- Novella, J. L., Boyer, F., Jochum, C., Jovenin, N., Morrone, I., Jolly, D., Bakchine, S., & Blanchard, F. 2006, "Health status in patients with Alzheimer's disease: an investigation of inter-rater agreement", *Qual Life Res.*, vol. 15, no. 5, pp. 811-819.
- O'Reilly, J., Lowson, K., Green, J., Young, J. B., & Forster, A. 2008, "Post-acute care for older people in community hospitals--a cost-effectiveness analysis within a multi-centre randomised controlled trial", *Age Ageing.*, vol. 37, no. 5, pp. 513-520.
- O'Reilly, J., Lowson, K., Young, J., Forster, A., Green, J., & Small, N. 2006, "A cost effectiveness analysis within a randomised controlled trial of post-acute care of older people in a community hospital", *BMJ.*, vol. 333, no. 7561, p. 228.
- O'Shea, E. 2004, "An economic and social evaluation of the Senior Help Line in Ireland", *Ageing & Society*, vol. 26, pp. 267-284.

- Oliver, D., Daly, F., Martin, F. C., & McMurdo, M. E. 2004, "Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review", *Age Ageing*, vol. 33, no. 2, pp. 122-130.
- Orme, J. G. & Reis, J. 1991, "Multiple regression with missing data", *Journal of Social Service Research*, vol. 15, no. 1/2, pp. 61-91.
- Ostbye, T., Tyas, S., McDowell, I., & Koval, J. 1997, "Reported activities of daily living: agreement between elderly subjects with and without dementia and their caregivers", *Age Ageing*, vol. 26, no. 2, pp. 99-106.
- Ouellet, M., Rivero-Arias, O., Gray, A., Wolstenholme, J., Rothwell, P. M., & Luengo-Fernandez, R. Mapping the Modified Ranking Scale (mRS) measurement into the generic EuroQol EQ-5D health outcome. Health Economics Study Group, University of East Anglia, January 2008.
- Paltiel, A. D., Fuhlbrigge, A. L., Kitch, B. T., Liljas, B., Weiss, S. T., Neumann, P. J., & Kuntz, K. M. 2001, "Cost-effectiveness of inhaled corticosteroids in adults with mild-to-moderate asthma: results from the asthma policy model", *J Allergy Clin Immunol*, vol. 108, no. 1, pp. 39-46.
- Parker, G., Bhakta, P., Katbamna, S., Lovett, C., Paisley, S., Parker, S., Phelps, K., Baker, R., Jagger, C., Lindesay, J., Shepperdson, B., & Wilson, A. 2000, "Best place of care for older people after acute and during subacute illness: a systematic review", *J Health Serv Res Policy*, vol. 5, no. 3, pp. 176-189.
- Parker, G. M., Katbamna, S., Bhakta, P., Phelps, K., Lovett, C., & Parker, S. 1999, *Best Place of Care for Older People after Acute and during Sub-acute Illness: a Systematic Review*, Nuffield Community Care Studies Unit, University of Leicester., Leicester.
- Parker, S. G., Peet, S. M., McPherson, A., Cannaby, A. M., Abrams, K., Baker, R., Wilson, A., Lindesay, J., Parker, G., & Jones, D. R. 2002, "A systematic review of discharge arrangements for older people", *Health Technology Assessment*, vol. 6, no. 4.
- Payakachat, N., Summers, K. H., Pleil, A. M., Murawski, M. M., Thomas, J., III, Jennings, K., & Anderson, J. G. 2009, "Predicting EQ-5D utility scores from the 25-item National Eye Institute Vision Function Questionnaire (NEI-VFQ 25) in patients with age-related macular degeneration", *Qual Life Res*, vol. 19.
- Peet, S., Phelps K., & Parker, G. 2002, *An evaluation of the effectiveness of admission avoidance, early discharge and community rehabilitation schemes in Leicestershire and Rutland*, Nuffield Community Care Studies Unit, University of Leicester, Leicester.
- Pegram, A. 1999, "What is case study research?", *Nurse Researcher*, vol. 7, no. 2, pp. 5-16.

Perry, L. & McLaren, S. 2004, "An exploration of nutrition and eating disabilities in relation to quality of life at 6 months post-stroke", *Health Soc.Care Community.*, vol. 12, no. 4, pp. 288-297.

Petch, A. 2003, *Intermediate Care: What Do We Know about Older People's Experiences?*, Joseph Rowntree Foundation, York.

Pickard, A. S., Wang, Z., Walton, S. M., & Lee, T. A. 2005, "Are decisions using cost-utility analyses robust to choice of SF-36/SF-12 preference-based algorithm?", *Health Qual Life Outcomes.*, vol. 3:11., p. 11.

Post, P. N., Stiggelbout, A. M., & Wakker, P. P. 2001, "The utility of health states after stroke: a systematic review of the literature", *Stroke.*, vol. 32, no. 6, pp. 1425-1429.

Potthoff, R. F., Tudor, G. E., Pieper, K. S., & Hasselblad, V. 2006, "Can one assess whether missing data are missing at random in medical studies?", *Stat.Methods Med Res.*, vol. 15, no. 3, pp. 213-234.

Powell, J. L. 1984, "Least absolute deviations estimation for the censored regression model", *Journal of Econometrics.*, vol. 25, pp. 303-325.

Power, C., Graham, H., Due, P., Hallqvist, J., Joung, I., Kuh, D., & Lynch, J. 2005, "The contribution of childhood and adult socioeconomic position to adult obesity and smoking behaviour: an international comparison", *Int.J Epidemiol.*, vol. 34, no. 2, pp. 335-344.

Raboud, J. M., Montaner, J. S., Thorne, A., Singer, J., & Schechter, M. T. 1996, "Impact of missing data due to dropouts on estimates of the treatment effect in a randomized trial of antiretroviral therapy for HIV-infected individuals. Canadian HIV Trials Network A002 Study Group", *J Acquir.Immune.Defic.Syindr.Hum.Retrovirol.*, vol. 12, no. 1, pp. 46-55.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. 2001, "A multivariate technique for multiply imputing missing values using a sequence of regression models", *Survey Methodology*, vol. 27, pp. 85-95.

Rait, G., Fletcher, A., Smeeth, L., Brayne, C., Stirling, S., Nunes, M., Breeze, E., Ng, E. S., Bulpitt, C. J., Jones, D., & Tulloch, A. J. 2005, "Prevalence of cognitive impairment: results from the MRC trial of assessment and management of older people in the community", *Age Ageing.*, vol. 34, no. 3, pp. 242-248.

Raymond, M. R. & Roberts, D. M. 1987, "A comparison of methods for treating incomplete data in selection research", *Educational and Psychological Measurement*, vol. 47, pp. 13-26.

Regen, E., Martin, G., Glasby, J., Hewitt, G., Nancarrow, S., & Parker, H. 2008, "Challenges, benefits and weaknesses of intermediate care: results from five UK case study sites", *Health Soc.Care Community.*, vol. 16, no. 6, pp. 629-637.

Revicki, D. A. 1992, "Relationship between health utility and psychometric health status measures", *Med Care.*, vol. 30, no. 5 Suppl, p. MS274-MS282.

- Revicki, D. A. & Kaplan, R. M. 1993, "Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life", *Qual.Life Res.*, vol. 2, no. 6, pp. 477-487.
- Revicki, D. A., Weinstein, M. C., & Alderman, M. 1992, *Health Utility and Health Status Outcomes of Antihypertensive Treatment*, Battelle Medical Technology Assessment and Policy Research Center, Washington, DC.
- Revicki, D. A., Wu, A. W., & Murray, M. I. 1995, "Change in clinical status, health status, and health utility outcomes in HIV-infected patients", *Med Care.*, vol. 33, no. 4 Suppl, p. AS173-AS182.
- Ribera, A., Permanyer-Miralda, G., Alonso, J., Cascant, P., Soriano, N., & Brotons, C. 2006, "Is psychometric scoring of the McNew Quality of Life after Myocardial Infarction questionnaire superior to the clinimetric scoring? A comparison of the two approaches", *Qual.Life Res.*, vol. 15, no. 3, pp. 357-365.
- Richardson, G., Griffiths, P., Wilson-Barnett, J., Spilsbury, K., & Batehup, L. 2001, "Economic evaluation of a nursing-led intermediate care unit", *International Journal of Technology Assessment in Health Care*, vol. 17, no. 3, pp. 442-450.
- Ridout, M. S. 1991, "Testing for random dropouts in repeated measurement data", *Biometrics.*, vol. 47, no. 4, pp. 1617-1619.
- Ritvo, P., Irvine, J., Naglie, G., Tomlinson, G., Bezjak, A., Matthew, A., Trachtenberg, J., & Krahn, M. 2005, "Reliability and validity of the PORPUS, a combined psychometric and utility-based quality-of-life instrument for prostate cancer", *J Clin.Epidemiol.*, vol. 58, no. 5, pp. 466-474.
- Roberts, K. 2002, "Exploring participation: older people on discharge from hospital", *J.Adv.Nurs.*, vol. 40, no. 4, pp. 413-420.
- Roberts, T., Henderson, J., Mugford, M., Bricker, L., Neilson, J., & Garcia, J. 2002, "Antenatal ultrasound screening for fetal abnormalities: a systematic review of studies of cost and cost effectiveness", *BJOG.*, vol. 109, no. 1, pp. 44-56.
- Roderick, P., Low, J., Day, R., Peasgood, T., Mullee, M. A., Turnbull, J. C., Villar, T., & Raftery, J. 2001, "Stroke rehabilitation after hospital discharge: a randomized trial comparing domiciliary and day-hospital care", *Age Ageing.*, vol. 30, no. 4, pp. 303-310.
- Roe, B. 2005, "Service development and evaluation in intermediate care," in *Intermediate and Continuing Care: Policy and Practice*, B. Roe & R. Beech, eds., Blackwell Publishing Ltd, Oxford, pp. 61-76.
- Roe, B., Daly, S., Shenton, G., & Lochhead, Y. 2003, "Development and evaluation of intermediate care", *J Clin.Nurs.*, vol. 12, no. 3, pp. 341-350.
- Rosenbaum, P. R. 2002, *Observational Studies*, 2nd edn, Springer Verlag, New York, NY.

- Rosenbaum, P. R. & Rubin, D. B. 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, vol. 70, pp. 41-55.
- Rosenbaum, P. R. & Rubin, D. B. 1984, "Difficulties with regression analyses of age-adjusted rates", *Biometrics*, vol. 40, no. 2, pp. 437-443.
- Rossi, P. H. & Freeman, H. E. 1993, *Evaluation: A Systematic Approach*, Sage Publications, Newbury Park, CA.
- Round, A., Crabb, T., Buckingham, K., Mejzner, R., Pearce, V., Ayres, R., Weeks, C., & Hamilton, W. 2004, "Six month outcomes after emergency admission of elderly patients to a community or a district general hospital", *Fam.Pract.*, vol. 21, no. 2, pp. 173-179.
- Royal Commission on Long Term Care for the Elderly 1999, *With Respect to Old Age: Long Term Care - Rights and Responsibilities*, The Stationery Office, London.
- Rubin, D. B. 1976, "Inference and missing data. ", *Biometrika*, vol. 61, pp. 581-592.
- Rubin, D. B. 1987, *Multiple Imputation for Nonresponse in Surveys*. . 160 John Wiley & Sons, New York.
- Saarni, S. I., Harkanen, T., Sintonen, H., Suvisaari, J., Koskinen, S., Aromaa, A., & Lonnqvist, J. 2006, "The impact of 29 chronic conditions on health-related quality of life: a general population survey in Finland using 15D and EQ-5D", *Qual.Life Res.*, vol. 15, no. 8, pp. 1403-1414.
- Sainsbury, A., Seebass, G., Bansal, A., & Young, J. B. 2005, "Reliability of the Barthel Index when used with older people", *Age Ageing*, vol. 34, no. 3, pp. 228-232.
- Sanderson, K., Andrews, G., Corry, J., & Lapsley, H. 2004, "Using the effect size to model change in preference values from descriptive health status", *Qual.Life Res.*, vol. 13, no. 7, pp. 1255-1264.
- Schafer, J. L. 1997, *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.
- Schafer, J. L. NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT, available from <http://www.stat.psu.edu/jls/misoftwa.html>. 1999. [Accessed 20 January 2008]
- Schafer, J. L. & Olsen, M. K. 1998, "Multiple imputation for multivariate missing-data problems: A data analyst's perspective", *Multivariate Behavioral Research*, vol. 33, no. 4, pp. 545-571.
- Schemper, M. 2003, "Predictive accuracy and explained variation", *Stat.Med.*, vol. 22, no. 14, pp. 2299-2308.
- Scuffham, P., Chaplin, S., & Legood, R. 2003, "Incidence and costs of unintentional falls in older people in the United Kingdom", *J.Epidemiol.Community Health*, vol. 57, no. 9, pp. 740-744.

- Segal, L., Day, S. E., Chapman, A. B., & Osborne, R. H. 2004, "Can we reduce disease burden from osteoarthritis?", *Med J Aust.*, vol. 180, no. 5 Suppl, p. S11-S17.
- Sengupta, N., Nichol, M. B., Wu, J., & Globe, D. 2004, "Mapping the SF-12 to the HUI3 and VAS in a managed care population", *Med Care.*, vol. 42, no. 9, pp. 927-937.
- Seshamani, M. & Gray, A. 2004a, "Ageing and health-care expenditure: the red herring argument revisited", *Health Econ.*, vol. 13, no. 4, pp. 303-314.
- Seshamani, M. & Gray, A. M. 2004b, "A longitudinal study of the effects of age and time to death on hospital costs", *J.Health Econ.*, vol. 23, no. 2, pp. 217-235.
- Shah, S., Vanclay, F., & Cooper, B. 1989, "Improving the sensitivity of the Barthel Index for stroke rehabilitation", *J Clin.Epidemiol.*, vol. 42, no. 8, pp. 703-709.
- Shao, J. & Zhong, B. 2003, "Last observation carry-forward and last observation analysis", *Stat.Med.*, vol. 23, pp. 3241-3244.
- Shaw, F. E., Bond, J., Richardson, D. A., Dawson, P., Steen, I. N., McKeith, I. G., & Kenny, R. A. 2003, "Multifactorial intervention after a fall in older people with cognitive impairment and dementia presenting to the accident and emergency department: randomised controlled trial", *BMJ.*, vol. 326, no. 7380, p. 73.
- Shepperd, S. & Iliffe, S. 2005, "Hospital at home versus in-patient hospital care", *Cochrane.Database.Syst.Rev.*, vol. 20; no. 3, p. CD000356.
- Sherbourne, D. C., Unutzer, J., Schoenbaum, M., Duan, N., Lenert, L. A., Sturm, R., & Wells, K. B. 2001, "Can utility-weighted health-related quality-of-life estimates capture health effects of quality improvement for depression?", *Med Care.*, vol. 39, no. 11, pp. 1246-1259.
- Shih, Y. C., Wang, X. S., Cantor, S. B., & Cleeland, C. S. 2006, "The association between symptom burdens and utility in Chinese cancer patients", *Qual.Life Res.*, vol. 15, no. 8, pp. 1427-1438.
- Shmueli, A. 1998, "The SF-36 profile and health-related quality of life: an interpretative analysis", *Qual.Life Res.*, vol. 7, no. 3, pp. 187-195.
- Shmueli, A. 1999, "Subjective health status and health values in the general population", *Med Decis.Making.*, vol. 19, no. 2, pp. 122-127.
- Shmueli, A. 2004, "The relationship between the visual analog scale and the SF-36 scales in the general population: an update", *Med Decis.Making.*, vol. 24, no. 1, pp. 61-63.
- Skinner, C. & Rao, J. N. K. 2002, "Jack-knife variance estimation for multivariate statistics under hot deck imputation from common donors", *Journal of Statistical Planning and Inference*, vol. 102, pp. 421-422.

Smith, A. E., Sim, J., Scarf, T., & Phillipson, C. 2004, "Determinants of quality of life amongst older people in deprived neighbourhoods", *Ageing & Society*, vol. 24, no. 5, pp. 793-814.

Spiers, N. A., Matthews, R. J., Barker, G. A., Jagger, C., & Hancock, R. M. 2005, "Standard of living in the retirement survey: a predictor of six-year functional limitation onset and mortality in women, but not in men from ages 55 - 69 years", *Disabil.Rehabil.*, vol. 27, no. 23, pp. 1415-1423.

Starkie, H. J., Briggs, A. H., & Chambers, M. G. Mapping the St George's Respiratory questionnaire to utility values using the EQ-5D. Health Economics Study Group, University of East Anglia, January 2008.

StataCorp L.P. Intercooled Stata 8.2 for Windows. 2004. TX, College Station.

Steiner, A. "Evidence and Evaluation Presentation to Kings Fund Conference 'Care closer to home 15 September, Ambassadors Hotel", in *Intermediate Care; A discussion Paper arising from the Kings Fund seminar held on 30th October 1996*, A. Steiner & B. Vaughan, eds., Kings Fund Publishing, London.

Steiner, A. 1997b, *Intermediate Care: A Conceptual Framework and Review of the Literature.*, Kings Fund, London.

Steiner, A. 2001, "Intermediate care--a good thing?", *Age Ageing.*, vol. 30 Suppl 3:33-9., pp. 33-39.

Steiner, A. & Vaughan, B. 1997, *Intermediate Care: A Discussion Paper arising from the Kings Fund Seminar held on 30th October 1996.*, Kings Fund, London..

Stevenson, J. & Spencer, L. 2002, *Developing Intermediate Care: A guide for Health and Social Services Professionals.*, Kings Fund, London.

Sullivan, P. W. & Ghushchyan, V. 2006a, "Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample", *Med Decis Making.*, vol. 26, no. 4, pp. 401-409.

Sullivan, P. W. & Ghushchyan, V. 2006b, "Preference-Based EQ-5D index scores for chronic conditions in the United States", *Med.Decis.Making.*, vol. 26, no. 4, pp. 410-420.

Thein, H. H., Krahn, M., Kaldor, J. M., & Dore, G. J. 2005, "Estimation of utilities for chronic hepatitis C from SF-36 scores", *Am J Gastroenterol.*, vol. 100, no. 3, pp. 643-651.

Thomas, K. J., MacPherson, H., Ratcliffe, J., Thorpe, L., Brazier, J., Campbell, M., Fitter, M., Roman, M., Walters, S., & Nicholl, J. P. 2005, "Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain", *Health Technol.Assess.*, vol. 9, no. 32, pp. iii-x, 1.

Tomassini, C. 2005, "The demographic characteristics of the oldest old in the United Kingdom", *Popul.Trends.* no. 120, pp. 15-22.

- Tomassini, C. 2006, "The oldest old in Great Britain: change over the last 20 years 149", *Popul.Trends.* no. 123, pp. 32-39.
- Torrance, G. W. 1982, "Preferences for health states: a review of measurement methods", *Mead Johnson Symp.Perinat.Dev.Med.* no. 20, pp. 37-45.
- Torrance, G. W. 1986, "Measurement of health state utilities for economic appraisal", *J Health Econ.*, vol. 5, no. 1, pp. 1-30.
- Torrance, G. W., Boyle, M. H., & Horwood, S. P. 1982, "Application of multi-attribute utility theory to measure social preferences for health states", *Oper.Res.*, vol. 30, no. 6, pp. 1043-1069.
- Torrance, G. W. & Feeny, D. 1989, "Utilities and quality-adjusted life years", *Int.J Technol.Assess.Health Care.*, vol. 5, no. 4, pp. 559-575.
- Tsevat, J., Solzan, J. G., Kuntz, K. M., Ragland, J., Currier, J. S., Sell, R. L., & Weinstein, M. C. 1996, "Health values of patients infected with human immunodeficiency virus. Relationship to mental health and physical functioning", *Med Care.*, vol. 34, no. 1, pp. 44-57.
- Tsevat, J., Solzan, J. G., & Tolan, K. K. Health status and utilities of HIV+ and primary care patients. *Clin.Res.* 40[569A]. 1992. [Abstract]
- Tsuchiya, A., Brazier, J., McColl, E., & Parkin, D. 2002, *Deriving preference-based single indices from non-preference based condition-specific instruments: converting AQLQ into EQ-5D indices. Sheffield Health Economics Group Discussion Paper 02/01*, School of Health and Related Research, Univeristy of Sheffield.
- United Nations 2000, *The World's Women 2000: Trends and Statistics*, United Nations Statistics Division.
- United Nations 2006, *World Population Prospects: The 2006 Revision*, United Nations Population Division, DESA, New York.
- Vach, W. & Blettner, M. 1999, "Logistic regression with incompletely observed categorical covariates -- investigating sensitivity against violation of the Missing at Random assumption", *Statistics in Medicine*, vol. 54, pp. 1315-1329.
- van den Bos, G. A. & Triemstra, A. H. 1999, "Quality of life as an instrument for need assessment and outcome assessment of health care in chronic patients", *Qual.Health Care.*, vol. 8, no. 4, pp. 247-252.
- van Doorslaer, E. & Jones, A. M. 2003, "Inequalities in self-reported health: validation of a new approach to measurement", *J Health Econ.*, vol. 22, no. 1, pp. 61-87.
- van Exel, N. J., Scholte op Reimer, W. J., & Koopmanschap, M. A. 2004, "Assessment of post-stroke quality of life in cost-effectiveness studies: the usefulness of the Barthel Index and the EuroQoL-5D", *Qual.Life Res.*, vol. 13, no. 2, pp. 427-433.

- Vaughan, B. & Lathlean, J. 1999, *Intermediate Care, Models in Practice.*, Kings Fund, London.
- Velikova, G., Stark, D., & Selby, P. 1999, "Quality of life instruments in oncology", *Eur.J Cancer.*, vol. 35, no. 11, pp. 1571-1580.
- Verbeke, G. & Molenberghs, G. 2000, *Linear Mixed Models for Longitudinal Data* Springer., New York.
- Vijverberg, W. 1987, "Non-Normality as Distributional Misspecification in Single-Equation Limited Dependent Variable Models.", *Oxford Bulletin of Economics and Statistics*, vol. 49, no. 4, pp. 417-430.
- Wade, D. T., Gage, H., Owen, C., Trend, P., Grossmith, C., & Kaye, J. 2003, "Multidisciplinary rehabilitation for people with Parkinson's disease: a randomised controlled study", *J Neurol.Neurosurg.Psychiatry.*, vol. 74, no. 2, pp. 158-162.
- Walker, L. & Jamrozik, K. 2005, "Effectiveness of screening for risk of medical emergencies in the elderly", *Age Ageing.*, vol. 34, no. 3, pp. 238-242.
- Walker, L., Jamrozik, K., Wingfield, D., & Lawley, G. 2005, "Increased use of emergency services by older people after health screening", *Age Ageing.*, vol. 34, no. 5, pp. 480-485.
- Walsh, B., Steiner, A., Pickering, R. M., & Ward-Basu, J. 2005, "Economic evaluation of nurse led intermediate care versus standard care for post-acute medical patients: cost minimisation analysis of data from a randomised controlled trial", *BMJ.*, vol. 330, no. 7493, p. 699.
- Walters, S. J. & Brazier, J. E. 2005, "Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D", *Qual Life Res.*, vol. 14, no. 6, pp. 1523-1532.
- Ward, S., Barnes, H., & Ward, R. 2005, "Evaluating a respiratory intermediate care team", *Nurs.Stand.*, vol. 20, no. 5, pp. 46-50.
- White, H. 1980, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity, in:", *Econometrica*, vol. 48, pp. 817-838.
- Wilkinson, P. R., Wolfe, C. D., Warburton, F. G., Rudd, A. G., Howard, R. S., Ross-Russell, R. W., & Beech, R. 1997, "Longer term quality of life and outcome in stroke patients: is the Barthel index alone an adequate measure of outcome?", *Qual.Health Care.*, vol. 6, no. 3, pp. 125-130.
- Wilson, A., Parker, H., Wynn, A., Jagger, C., Spiers, N., Jones, J., & Parker, G. 1999, "Randomised controlled trial of effectiveness of Leicester hospital at home scheme compared with hospital care", *BMJ.*, vol. 319, no. 7224, pp. 1542-1546.
- Wilson, A., Parker, H., Wynn, A., & Spiers, N. 2003, "Performance of hospital-at-home after a randomised controlled trial", *Journal of Health Services Research Policy*, vol. 8, no. 3, pp. 160-164.

Wilson-Barnett, J., Davis, T., Griffiths, P., Harris, R., & Heard, S. 2001, "Evaluating nurse-led inpatient units", *Nurs.Stand.*, vol. 15, no. 37, pp. 40-42.

Wise, D. A. 1997, "Retirement against the demographic trend: more older people living longer, working less, and saving less", *Demography.*, vol. 34, no. 1, pp. 83-95.

Wolfe, C. D., Taub, N. A., Woodrow, E. J., & Burney, P. G. 1991, "Assessment of scales of disability and handicap for stroke patients", *Stroke.*, vol. 22, no. 10, pp. 1242-1244.

Wooldridge, J. M. 2002, *Econometric Analysis of Cross Section and Panel Data* The MIT Press, Massachusetts.

Wothke, W. 1998, "Longitudinal and multi-group modeling with missing data. " in *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples*, T. D. Little, K. U. Schnabel, & J. Baumert, eds., Lawrence Erlbaum Publishers., Mahwah, NJ.

Wu.E, Mulani, P., Farrell, M. H., & Sleep, D. Mapping Fact-P and Eortc Qlq-C30 to the Eq 5D Health Utility in Metastasis Hormone-Refractory Prostate Cancer Patients. *Value Health.* 9, A114. 2006. [Abstract]

Yang, M., Dubois, D., Kosinski, M., Sun, X., & Gajria, K. 2007, "Mapping MOS Sleep Scale scores to SF-6D utility index", *Curr.Med Res Opin.*, vol. 23, no. 9, pp. 2269-2282.

Yohannes, A., Roomi, J., Waters, K., & Connolly, M. 1998, "A comparison of the Barthel index and Nottingham extended activities of daily living scale in the assessment of disability in chronic airflow limitation in old age", *Age and Ageing*, vol. 27, no. 3, pp. 369-374.

Young, J. 2002, "The evidence base for intermediate care", *Geriatric Medicine*, vol. 10, pp. 11-14.

Young, J., Forster, A., & Green, J. 2003, "An estimate of post-acute intermediate care need in an elderly care department for older people", *Health Soc.Care Community.*, vol. 11, no. 3, pp. 229-231.

Young, J., Forster, A., Green, J., & Bogle, S. 2007, "Post-acute transfer of older people to intermediate care services: the sooner the better?", *Age Ageing.*, vol. 36, no. 5, pp. 589-592.

Young, J., Robinson, M., Chell, S., Sanderson, D., Chaplin, S., Burns, E., & Fear, J. 2005a, "A prospective baseline study of frail older people before the introduction of an intermediate care service", *Health Soc.Care Community.*, vol. 13, no. 4, pp. 307-312.

Young, J. B., Robinson, M., Chell, S., Sanderson, D., Chaplin, S., Burns, E., & Fear, J. 2005b, "A whole system study of intermediate care services for older people", *Age Ageing.*, vol. 34, no. 6, pp. 577-583.

Yu, A. P., Yu, Y. F., & Nichol, M. B. 2005, "Bridging beyond regression-artificial neural network in mapping from SF-36 to health utility", *Value Health.*, vol. 8, p. 384.

Zheng, B. & Agresti, A. 2000, "Summarizing the predictive power of a generalized linear model", *Stat.Med.*, vol. 19, no. 13, pp. 1771-1781.